

**SIXTH FRAMEWORK PROGRAMME**  
**PRIORITY 2**  
**INFORMATION SOCIETY TECHNOLOGIES**



**SIXTH FRAMEWORK  
PROGRAMME**

**FLOSSWORLD**

**Free/Libre and Open Source Software: Worldwide  
Impact Study**



Track 2 Study Report – International Study Report  
(Referred to as D30a in the work packages description in the proposal)

Project Reference: 015722

Kind of Project: Specific Support Action

Start Date: May 2005

End Date: June 2007



UNITED NATIONS  
UNIVERSITY

**UNU-MERIT**



Universidad  
Rey Juan Carlos

## **Authors and contributors**

### *Coordination, Drafting and Edition*

Jesús M. González-Barahona, GSyC/LibreSoft, Universidad Rey Juan Carlos

Teófilo Romera, GSyC/Libresoft, Universidad Rey Juan Carlos

Daniel Izquierdo-Cortázar, GSyC/Libresoft, Universidad Rey Juan Carlos

Álvaro del Castillo, GSyC/Libresoft, Universidad Rey Juan Carlos

### *Reviewing*

Gregorio Robles, GSyC/Libresoft, Universidad Rey Juan Carlos

*Details about contributors in the areas of retrieval process, graphics, local support and others for the regional studies on which this report is based are available in the corresponding country reports, distributed by the FLOSSWorld project*

## **Copyright**

©2007 GSyC/LibreSoft

Some rights reserved. This report is distributed under the Creative Commons Attribution-ShareAlike 3.0 licence, available in <http://creativecommons.org/licenses/by-sa/3.0>

This report is available in <http://flossworld.org>

This report has been funded by the European Commission under contract number FP6-IST-015722.

## **Disclaimer**

The opinions expressed in this Study are those of the authors and do not necessarily reflect the views of the European Commission. Contract FP6-IST-015722

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Remarkable results</b>	<b>2</b>
<b>3</b>	<b>Methodology details</b>	<b>4</b>
3.1	Studies performed . . . . .	4
3.2	Data Sources . . . . .	5
3.3	Methodology . . . . .	6
3.4	Tools . . . . .	6
<b>4</b>	<b>Inter-regional differences</b>	<b>7</b>
4.1	Regional communities . . . . .	7
4.2	Number of developers per region . . . . .	8
4.3	Correlation of number of developers and other parameters . . . . .	10
4.4	Distribution of developers by time zone . . . . .	11
4.5	Distribution of authorship among the several countries . . . . .	12
<b>5</b>	<b>Regional results</b>	<b>13</b>
5.1	China . . . . .	13
5.1.1	Forges analysed . . . . .	13
5.1.2	Programming languages . . . . .	16
5.1.3	Authorship data . . . . .	17
5.1.4	Other data . . . . .	18
5.2	India . . . . .	18
5.2.1	Forges analysed . . . . .	19
5.2.2	Programming languages . . . . .	21
5.2.3	Authorship . . . . .	22
5.2.4	Other data . . . . .	23
5.3	South Africa . . . . .	23
5.3.1	Forges analysed . . . . .	24
5.3.2	Programming Languages . . . . .	26
5.3.3	Authorship . . . . .	27
5.3.4	Other data . . . . .	28
5.4	Brazil . . . . .	28
5.4.1	Forges analysed . . . . .	29
5.4.2	Programming Languages . . . . .	30
5.4.3	Authorship . . . . .	31
5.4.4	Other data . . . . .	32
5.5	Argentina . . . . .	32

5.5.1	Forges analysed . . . . .	33
5.5.2	Programming Languages . . . . .	35
5.5.3	Authorship . . . . .	36
5.5.4	Other data . . . . .	36
5.6	Malaysia . . . . .	37
5.6.1	Forges analysed . . . . .	37
5.6.2	Programming Languages . . . . .	38
5.6.3	Authorship . . . . .	39
5.6.4	Other data . . . . .	40
5.7	Croatia . . . . .	40
5.7.1	Forges analysed . . . . .	41
5.7.2	Programming languages . . . . .	42
5.7.3	Authorship . . . . .	43
5.7.4	Other data . . . . .	43
5.8	Bulgaria . . . . .	44
5.8.1	Forges analysed . . . . .	44
5.8.2	Programming languages . . . . .	45
5.8.3	Other data . . . . .	46

# List of Tables

4.1	Copyright owners found in software releases in forges . . . . .	12
5.1	Registered users and projects in Chinese forges (data 16th April 2007). SourceForge has been included for completeness (data June, 2006). . . . .	14
5.2	Information sources that could be extracted from Chinese forges (April-May 2007). . . . .	14
5.3	Programming languages used in Cosoft forge . . . . .	16
5.4	Type of authorship in Chinese forges . . . . .	17
5.5	Other data collected . . . . .	18
5.6	Registered users and projects in Indian forges, including SourceForge . . . . .	19
5.7	Information sources that could be studied in Indian forges (April 2007). . . . .	19
5.8	Programming Languages used in Sarovar forge . . . . .	21
5.9	Type of authorship in Indian forges . . . . .	22
5.10	Other data collected . . . . .	23
5.11	South African forges (on 16th of April of 2007) . . . . .	24
5.12	South African forges (on April-June of 2007). SourceForge has been included for completeness (data June, 2006) . . . . .	24
5.13	Programming languages used in Avoir forge . . . . .	26
5.14	Type of authorship in South African forges . . . . .	27
5.15	Other data collected . . . . .	28
5.16	Registered users and projects in Brazilian forges (30th of May 2007). SourceForge has been included for completeness (data June, 2006). . . . .	29
5.17	Information sources that could be extracted from Brazilian forges (April 2007). . . . .	30
5.18	Programming Languages used in CodigoLivre forge . . . . .	30
5.19	Type of authorship in Brazilian forges . . . . .	31
5.20	Other data collected . . . . .	32
5.21	Registered users and projects in Argentinian forges (April 2007). SourceForge has been included for completeness (data June, 2006). . . . .	33
5.22	Information sources that could be extracted from Argentinian forges (April-May 2007). . . . .	33
5.23	General Language results . . . . .	35
5.24	Type of authorship in Argentinian forges . . . . .	36
5.25	Other data collected . . . . .	36
5.26	Registered users and projects in Malaysian forges (data 16th April-May 2007). SourceForge has been included for completeness (data June, 2006). . . . .	37
5.27	Information sources that could be extracted from Malaysian forges (April-May 2007). . . . .	38
5.28	Programming languages used in <i>Osc Knowledge Bank</i> . . . . .	38
5.29	Type of authorship in Malaysian forges . . . . .	39
5.30	Other data collected . . . . .	40
5.31	Registered users and projects in Croatian forges (data April-May 2007). SourceForge has been included for completeness (data June, 2006). . . . .	41
5.32	Information sources that could be extracted from Croatian forges (April-May 2007). . . . .	41

5.33	General Language results . . . . .	42
5.34	Type of authorship in Croatian forges . . . . .	43
5.35	Other data collected . . . . .	43
5.36	Registered users and projects in Bulgaria forges (data 24th April 2007). SourceForge has been included for completeness (data June, 2006). . . . .	44
5.37	Information sources that could be extracted from Bulgarian forges (April-May 2007). . .	44
5.38	Programming languages used in Openfmi forge . . . . .	46
5.39	Other data collected . . . . .	47

# List of Figures

4.1	Information about regional libre software community indicators, as found by partners in the FLOSSWorld project . . . . .	8
4.2	New accounts in SourceForge, per country, per year (estimated) . . . . .	9
4.3	Accounts in SourceForge with CVS/SVN activity, per country (estimated), ca. 2006 . . . . .	9
4.4	Total number of accounts in SourceForge, per region, evolution (estimated) . . . . .	10
4.5	Accounts in SourceForge coloured in relationship with country population (estimated), ca. 2006 . . . . .	10
4.6	Accounts in SourceForge coloured in relationship with Internet usage (estimated), ca. 2006 . . . . .	11
4.7	Accounts in SourceForge coloured in relationship with GDP/capita (estimated), ca. 2006 . . . . .	11
4.8	Time zones of posters to project mailing lists, ca. 2007 . . . . .	12
5.1	SCM repositories, mailing lists and software releases found in forges . . . . .	15
5.2	Committers per forge (estimated, in the case of SourceForge) . . . . .	15
5.3	Commits per forge (estimated, in the case of SourceForge) . . . . .	16
5.4	Programming languages used in Cosoft forge . . . . .	17
5.5	Authorship data (in software releases). . . . .	18
5.6	SCM repositories, mailing lists and software releases found in forges . . . . .	20
5.7	Committers in some selected projects . . . . .	20
5.8	Commits in some selected projects . . . . .	21
5.9	Programming Languages used in Sarovar forge . . . . .	22
5.10	Authorship data (in software releases). . . . .	23
5.11	SCM repositories, mailing lists and software releases found in forges . . . . .	25
5.12	Committers per forge . . . . .	25
5.13	Commits per forge . . . . .	26
5.14	Programming languages used in Avoir forge . . . . .	27
5.15	Authorship data (in software releases). . . . .	28
5.16	Projects and users per forge . . . . .	29
5.17	Programming languages used in <i>CodigoLivre</i> forge . . . . .	31
5.18	Authorship data (in software releases). . . . .	32
5.19	SCM repositories, mailing lists and software releases found in forges . . . . .	34
5.20	Committers per forge . . . . .	34
5.21	Commits per forge . . . . .	35
5.22	Information of languages used in the develop of projects . . . . .	36
5.23	Programming languages used in <i>Ossc Knowledge Bank</i> . . . . .	39
5.24	Authorship data (in software releases). . . . .	40
5.25	SCM repositories, mailing lists and software releases found in forges . . . . .	42
5.26	Programming languages in <i>Linx.hr</i> . . . . .	43
5.27	Committers per forge . . . . .	45
5.28	Commits per forge . . . . .	45
5.29	Programming languages used in Openfmi forge . . . . .	46

# Chapter 1

## Introduction

As a part of the FLOSSWorld project, a quantitative study of several aspects of libre software development and libre software communities in eight target regions has been performed. Those regions are: China, India, South Africa, Brazil, Argentina, Malaysia, Croatia and Bulgaria. For each of them, an exhaustive search of indicators related to their libre software communities (mainly by the local partners of the project), and an in-depth analysis of the libre software projects hosted in them have been performed. As a result of these studies, this quantitative part of the project has generated:

- Eight regional reports (one per target country)
- One methodology report, showing and discussing details about the data sources, the methodologies used for data retrieval and analysis, and the tools used for those tasks, when relevant.
- This summary report, which includes the main results, is self-contained and has been designed to be useful if read in isolation. It is specially indicated for those willing to know with certain details the main findings of the project.

In order to fulfil its objectives, this report provides information not just about the main results of the project, but also about how they were obtained, and what data sources were used. Its structure is as follows:

- Next chapter (chapter 2) shows the most remarkable results of this part of the project. This chapter is self-contained, and is probably the first one that should be read in order to get an idea of the findings of the project.
- Chapter 3 shows a summary of the studies performed in this part of the FLOSSWorld project, the methodology used to find, retrieve and analyse the data used in the studies. Details about this methodology can be found in the separate “Methodology Report”.
- Chapter 4 addresses the inter-regional differences found between target regions, and between them and the global libre software community.
- Chapter 5 details the main findings of the studies performed on the data of each of the eight target regions. Details of this information can be found in the eight separate country reports.

This report was written during the first half of 2007. Most of the data used in it is from early 2007 or 2006 (the details can be found in the Methodology report).

## Chapter 2

# Remarkable results

Some of the most remarkable findings of the project, in the area of quantitative information about libre software development in the target regions, are:

- Local and regional libre software communities do exist, and they maintain their own infrastructure for software development. This result was clearly expected by the project, but has been backed with detailed data. Regional communities have been characterised, finding how diverse they are, and providing a first estimation about their size in the target regions (see details in chapter 5 about data per country, including for instance, the number of people involved in development, the number of projects, the kind of forges in each country, etc).
- However, these local and regional communities are small in size when compared with the global libre software community. This result was expected, but the report has been able of showing to which extent most libre software development is really a global activity. It is also remarkable that many of the target countries (such as China, Brazil or India) are among those more likely to have large regional development communities. This means that this result can be considered as meaningful not only for the target countries, but in general.
- Some global projects have been found in regional forges in the target countries, specially in India. This is a clear indicator that the infrastructure for global software development is not only provided by developed countries, and that the results of the report have to be considered with care, since some of the development identified as regional could be in fact a part of global libre software development.
- There is a lot of variety in the kind of web sites for supporting regional libre software development. These sites, usually called “forges”, are not always run by GForge-like<sup>1</sup>, systems, although that is the trend, as is the case worldwide. What is maybe more surprising is that they are in several cases not even “true” development sites, but more like repositories of software developed elsewhere. This means that, contrary to the main trend in the global libre software community, many regional developments produce public versions of their software, but use little of the usual infrastructure for developing in the open source way (with public source code management systems, mailing lists and bug tracking systems, for instance).
- Regional communities are usually focused on developments that are local, meant for the use of the regional communities themselves. Except some specific cases (India is the most notable, Brazil is also an exception to some extent), the projects and products are not known neither used in other regions, neither are they known to the global software community. Developers also seem

---

<sup>1</sup>GForge is a software for maintaining development forges, produced after a fork of the SourceForge software

to be from the involved region, with little participation of foreign developers (again, India is an exception to this).

- A great share of the projects carried on by regional communities are localisation of global packages. Brazil (which produces many “original” packages) and India (where many projects seem to be no much interested in localisation) are the most clear exceptions.
- Local languages are usually used for communication in these regional communities. This is almost an absolute fact in countries such as China, Argentina or Brazil. However, there are also some exceptions: English is almost the only language found in India, and very common in South Africa. Although both countries have a large fraction of their population being fluent in English (which explains the fact), it is interesting how local languages (specially in the case of India) are clearly under-represented in libre software development. Language also explains why the community in India, and with less intensity in South Africa, are clearly more interrelated with the global libre software development community than other target regions. Language also helps to explain the relative isolation of a large fraction of developers in regional communities: if they are not fluent in English, the *lingua franca* of the global libre software community, they are somehow restricted to work with their regional fellows who speak the same language.
- Scripting languages (Perl, Python, PHP) are clearly overrepresented in almost all regions (and even more clearly if localisations of Linux-based distributions are excluded from the analysis). This is curious, and could imply that not all programming technologies are equally interesting for developers in the target regions, or that they are not equally understandable or adaptable to local needs.

Many other quantitative results have been produced by the FLOSSWorld project. They are detailed in the rest of this report, in the regional reports, and in other reports published by the project.

## Chapter 3

# Methodology details

This chapter presents the details of the methodology used to perform the studies supporting the data presented in the report. These details are convenient for understanding the data presented in the next chapters, and to have some context to interpret their significance. In fact, not only the methodology is presented in the following sections, which also shows: a summary of the studies performed (including some quantitative data on the data sources), a complete description of the data sources used, the methodology for analysing each of them and a summary of the tools that helped to complete the studies.

### 3.1 Studies performed

The most relevant studies, whose results are presented in this report, are:

- Study of some metrics that can help to characterise the libre software community (and developer community) of each target region. For this study, some representative parameters (such as publications related to libre software, Linux user groups, or libre software development communities) have been identified. Partners in each of the target regions have conducted an extensive search for data to provide an accurate estimation of those parameters.
- Study on the SourceForge user database<sup>1</sup>, which includes information about mail addresses and time zones for registered users. This information is used to estimate the geographical origin of developers.
- Study of the mailing lists of several large global libre software projects. The mail addresses and time zones of messages are also used to estimate the origin of developers.
- Study of forges found in the target regions. In fact, this has been the largest part of the research performed to complete this report, and has included the analysis of public information about registered users, source code management repositories (CVS and Subversion), release files with source code, and mailing list archives. The information obtained from these studies has been used to estimate developer and project population in target regions, quantity of source code developed, etc.

The study targeted eight developing countries, focusing on evidences of local libre software development. SourceForge, the largest forge in the world, has been studied as a proxy of the global libre software development community, which provides a good contrast for all the regional data.

---

<sup>1</sup>This database is provided to researchers by by University of Notre Dame, see details at <http://www.nd.edu/~oss/Data/data.html>

Regarding to regional forges, 13 have been studied and they are the base for each regional study. In these forges, 655 committers (developers who use the source control management systems) have been detected and 428,453 commits have been counted working on 2,370 projects. In contrast, 1111 developers whose nationality is some of the regional studies have been estimated in SourceForge SCM repositories (for a total of about 30,000 committers), and they have made 391,714 commits working on 852 projects. Again in local forges, 370 software releases were found and could be analysed from local forges, amounting to 8,905,068 lines of code, and 284 mailing lists archives were analysed, having found 3,173 different email addresses and 11,120 messages.

Finally, every partner of the FLOSSWorld project provided information related to communities, LUGs and other forums of libre software interest. Specifically, our partners found 55 communities, 306 LUGs, 142 projects, 11 platforms, 7 elements set in media area and 303 developers.

## 3.2 Data Sources

The open nature of libre software projects usually implies that they offer not only the source code publicly on the Internet, but many other information sources. Although this information is usually targeted to lower the barrier of entry to new developers, so that they can read in the archives discussions and decisions that have happened, it also serves as a basis for the quantitative approach that we intend for the analysis of libre software projects.

In the context of the FLOSSWorld project, this means that we first have to identify projects from the countries under study. Here is where the local FLOSSWorld partners have provided the necessary information hinting to specific projects and repositories from their countries.

We have therefore differentiated between primary and secondary data sources. Primary data sources are the ones that have been used to identify projects and other libre software related information in the specific countries. This has been done in strong coordination with the local FLOSSWorld partners in each country. Secondary data sources are the ones that we actually analyse with the help of specific tools. Secondary data sources are obtained by further investigating the primary data sources.

### *Primary data sources:*

1. Survey (with collaboration of partners)
2. Internet Searches

### *Secondary data sources:*

1. Collaborative Development Environments (CDE or forges)
2. Repositories (Code Management Systems)
3. Mailing Lists
4. Source code releases

For each secondary data source we have conceived a download process. To identify the secondary data sources we have processed the data provided by the local FLOSSWorld partners by means of a survey. The information provided by local partners has been completed with Internet searches and with suggestions from experts in the libre software area.

Once we have the primary data sources, they have to be studied in detail in order to obtain information about secondary data sources, which are the ones that we can analyse semi-automatically by means of software tools. This is the case for instance of the platforms provided by the local FLOSSWorld partners; a spidering tool has been used to obtain a list of source code management repositories, mailing list archives and releases for projects hosted in them.

### 3.3 Methodology

The methodology consists of the steps that have been followed to download the data from the various data sources, how we have analysed these data and, finally, what other procedures have been applied to the process.

These steps are as follows<sup>2</sup>:

1. Survey (with collaboration of partners): The survey has been conceived to obtain information from local FLOSSWorld partners about libre software projects and libre software-related associations and media in their countries. It consists of a web-based platform where data can be introduced using forms.  
The information provided is divided in communities, developers, Linux user groups, media, platforms and projects.
2. Forges: In forges there are different data sources which can be analysed. We focused in repositories, mailing lists and releases, however, more information can be retrieved from forums, news and other kind of communication among users.
3. Analysis of global forges: This analysis has been done on SourceForge, the largest forge in the world. SourceForge can be considered a global forge as, although located in the United States, developers from all over the world participate in projects hosted there. We have applied a methodology that allows to infer the nationality of the SourceForge users based on their e-mail addresses and the time zone they specify when they register. However, in many cases the identification of the country of origin is not so simple. The reason for this is basically because we have to face incomplete information.<sup>3</sup>

### 3.4 Tools

The GSyC/LibreSoft group at the Universidad Rey Juan Carlos has got a collection of analysis tools that has been used for the analysis performed for the FLOSSWorld project. Some scripts have been created in order to automate the process of information retrieval.

1. ForgeSpider: It is a tool that extracts specific information out of GForge based forges. It retrieves information such as: Project names, repository's URLs, mailing lists archives and releases' URLs.
2. CVSAlyY: It is a tool that extracts statistical information out of CVS (and since recently, out of Subversion too) repository logs and transforms it in database SQL formats.
3. MailingListStats: It is a tool for mapping mbox files of any mailing list to a database.
4. PyTernity: It is a statistical analysis tool for building and analysing distributions of ownership/contribution data from software source packages. It has been primarily designed to study the patterns in contributions from developers working on libre software projects.
5. SLOCCount: It is a suite of programs for counting physical source lines of code (SLOC) in possibly large software systems. The tool can count physical SLOC for a wide number of languages; can take a large set of files and automatically categorise their types using a number of different heuristics; and also, comes with analysis tools.

---

<sup>2</sup>In order to obtain detailed information about methodology, please access to the *Methodology Report*

<sup>3</sup>For specific information about the whole process, please access to *Methodology Report - Chapter Methodology - section Analysis of global forges*

## Chapter 4

# Inter-regional differences

The conclusions shown in this chapter were obtained by comparing the studies on each of the target regions, and by comparing them with two analysis on the global software community that have also been performed in part in the context of the FLOSSWorld project: one based on a geographical analysis of data from SourceForge, the largest libre software forge worldwide<sup>1</sup>; and another based on country-analysis of mailing lists of some global libre software projects.

### 4.1 Regional communities

The different studies performed by the FLOSSWorld project have clearly shown that there are rich regional libre software development communities. They are also quite different from country to country in volume, degree of integration with the global libre software community, cultural practices and infrastructure at their disposal.

A first estimator of the communities per country is shown in figure 4.1, which depicts the data found by FLOSSWorld partners in their own regions, in several areas:

---

<sup>1</sup>Some parts of this study, including detailed information about methodology have been published as “Geographic Location of Developers at SourceForge”, by Gregorio Robles and Jesus M. Gonzalez-Barahona, in the Proceedings of the Third International Workshop on Mining Software Repositories, Co-located with the International Congress on Software Engineering (Shanghai, China, May 2006).

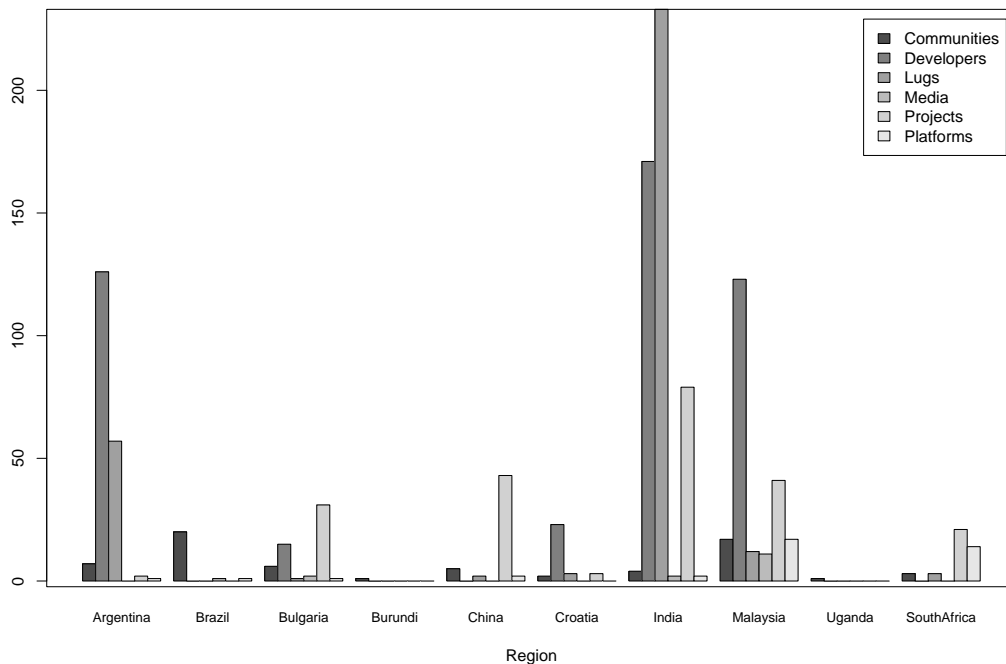


Figure 4.1: Information about regional libre software community indicators, as found by partners in the FLOSSWorld project

- **Communities:** virtual groups of users interested in libre software, usually meeting and sharing information at a web site (which is identified in the corresponding FLOSSWorld regional report).
- **LUGs:** Linux User Group, usually with a physical address, and also usually corresponding to a group or people meeting in person with some periodicity (although some of them are also mainly organised around web sites).
- **Media:** any kind of journal or magazine related to libre software.
- **Platforms:** web sites which provide some support to libre software development, usually development forges.

Since not all partners had the same facilities for getting the data (in some countries there are detailed and up-to-date catalogues of some of these items, for instance, while in others the community is fractioned and finding information about it as a whole is difficult), the information about all the countries has different degrees of reliability. But in any case, it shows both that several large regional communities exist, and that their characteristics are quite different from country to country.

## 4.2 Number of developers per region

In the FLOSSWorld project the number of developers per region has been estimated from several points of view. Figure 4.2 shows the estimation of new developers (accounts) in SourceForge, per country, for each of the target countries. Data is shown as new developers per year, which helps to appreciate the evolution of the situation in the different regions.

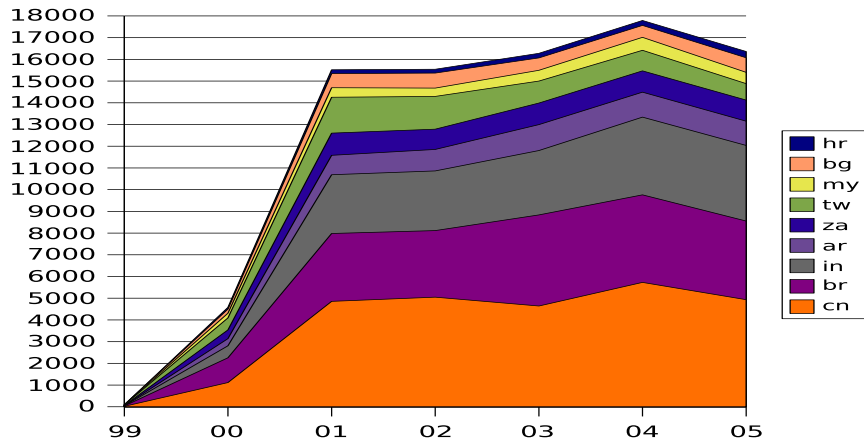


Figure 4.2: New accounts in SourceForge, per country, per year (estimated)

However, the number (and relative shares) of developers is different if we analyse it focusing on the most active developers (in the case of our study, those with activity in the source code management systems in SourceForge), as is shown in figure 4.3<sup>2</sup>.

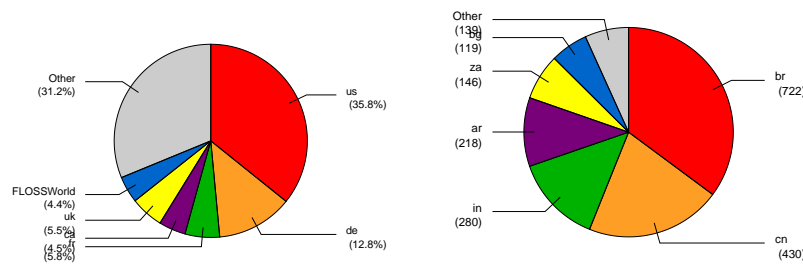


Figure 4.3: Accounts in SourceForge with CVS/SVN activity, per country (estimated), ca. 2006

The evolution of developers in the main areas of the world is also interesting, showing how developing countries are gaining a larger share of accounts in SourceForge with time (see figure 4.4). This probably means that the fraction of developers in the global libre software development community is also growing with time. When considering also the size of the local communities (not considered in the SourceForge graph), probably the fraction represented by developing countries in the worldwide community (including those local and regional communities) is even larger. However, North America and Europe still represent a clearly predominant share of libre software developers.

<sup>2</sup>These figures for the origin of developers active in SCM repositories are estimations, based on time zones and email addresses, and are therefore different than those presented in the regional reports, and in chapter 5, which for SourceForge developers active in SCM consider only email addresses, therefore missing those with generic ones (such as .com or .org). See details in the Methodology Report. This has been done to have data more comparable with countries with high involvement in the global libre software community.

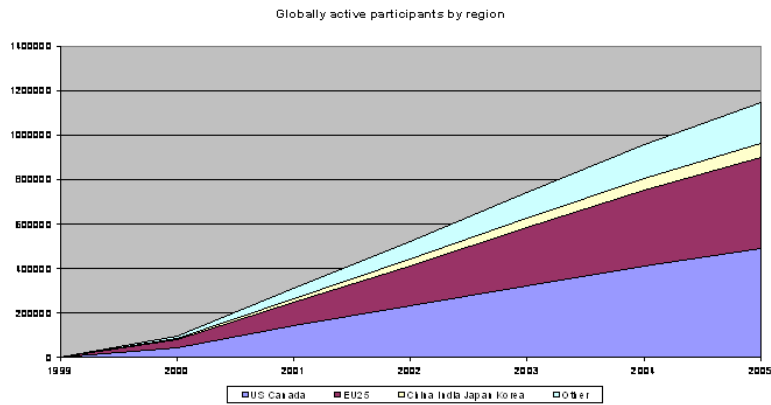


Figure 4.4: Total number of accounts in SourceForge, per region, evolution (estimated)

### 4.3 Correlation of number of developers and other parameters

Several parameters have been correlated with the number of estimated developers per country, worldwide, with the idea of finding good estimators for this number of developers. Figures 4.5, 4.6 and 4.7 show some coloured maps showing these correlations for population, Internet penetration and GDP per capita, which have been some of the most accurate ones. Of those, it can be observed that the most precise is GDP per capita (despite the difficulty to interpret that relationship, given the large differences in population from country to country). In all maps, most of Africa, Middle East, and some parts of South America appear in light colours, which mean bad correlation. This is due to the extreme low number of developers in those countries, which is an anomaly in itself (and difficult to explain, since it includes countries with very different populations and wealths).

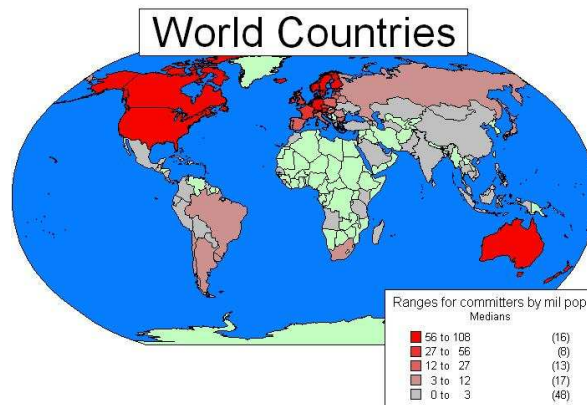


Figure 4.5: Accounts in SourceForge coloured in relationship with country population (estimated), ca. 2006

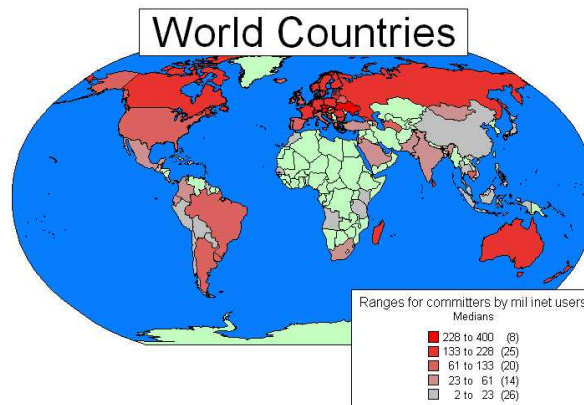


Figure 4.6: Accounts in SourceForge coloured in relationship with Internet usage (estimated), ca. 2006

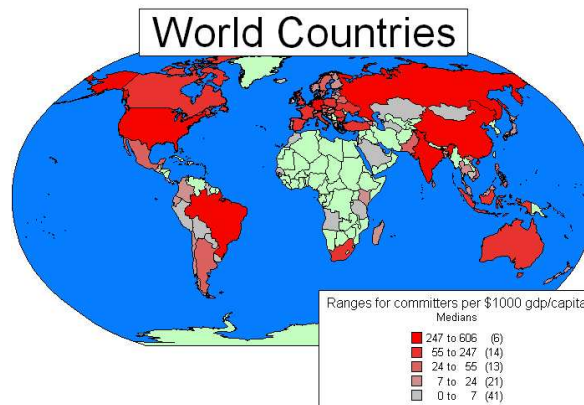


Figure 4.7: Accounts in SourceForge coloured in relationship with GDP/capita (estimated), ca. 2006

#### 4.4 Distribution of developers by time zone

Based on the analysis of mailing list archives of several large global libre software projects, the number of developers by time zone has been estimated (see figure 4.8). Although usually many countries share the same time zone, this information is useful for several reasons:

- In some cases, one country can be clearly identified as the main (and in some cases almost single) contributor for one time zone. This is for instance the case of India or the West Coast of USA.
- In some cases, individual countries cannot be identified as main contributors, but large regional areas can. For instance, time zones from -4 to -9 clearly correspond to the Americas, while -1 to 3 correspond to Europe and Africa (and given the very low number of developers found in Africa in other studies, it could be attributed almost in exclusive to Europe).
- In some cases, even intra-country differences can be found. For instance, in the Americas time zones, the impact of USA is clearly larger than any other country, and the differences between time zones in that area can be explained mainly in terms of distribution of developers within USA. From that point of view, the peaks at -8 and -9 correspond to the West Coast, while that at -5 corresponds mainly to the East Coast.

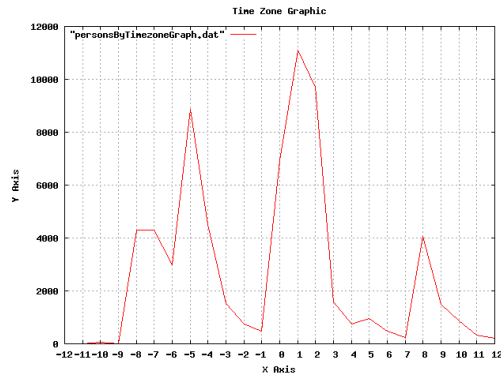


Figure 4.8: Time zones of posters to project mailing lists, ca. 2007

## 4.5 Distribution of authorship among the several countries

Copyright notices in the source code of software releases found in all analysed forges were studied, with the aim of identifying copyright holders (which are usually a good proxy for authorship). The results of this study are found in table 4.1 (only those countries where enough software releases were found are included in the table).

Some interesting patterns can be observed in it. China can be highlighted because of the importance of the number of companies found (90), while on the contrary Argentina does not show evidence of any company claiming copyright on the software studied.

Type	AR	CN	MY	ZA	IN	BR
<b>Ind. authors</b>	327	267	86	52	42	26
<b>Teams</b>	0	22	6	2	3	5
<b>Universities</b>	0	17	3	4	4	1
<b>Companies</b>	0	90	9	9	3	9
<b>Foundations</b>	0	5	1	1	1	1
<b>Unknown</b>	49	11	3	3	0	4

Table 4.1: Copyright owners found in software releases in forges

# Chapter 5

## Regional results

This chapter shows the most remarkable results of the community and software development studies, organized by studied region. The regional reports produced by the FLOSSMETRICS project show more detail on each of these regions.

As is explained with more detail in chapter 3, the results are based on the analysis performed on some quantitative aspects of libre (free, open source) software in the countries studied in the FLOSSWorld project. The sources of the analysis have been an exhaustive data retrieval of several facts related to libre software in each country (such as a list of Linux user groups, magazines focused on libre software, etc.) and a detailed, quantitative data mining of several sites hosting libre software development (forges).

In this chapter, each section corresponds to the main results of each of the studied countries.

### 5.1 China

Besides SourceForge (the largest forge in the world, for which the Chinese participation has been estimated), this study has analysed five Chinese forges, of which Cosoft, OSS and OSDN are the largest ones. Cosoft is the largest forge in China with more than 45,000 registered users and more than 1,450 registered projects, while OSS is the next one by number of registered users (with more than 6,500), and OSDN by number of registered projects (187). These figures clearly show the big gap between Cosoft and the other forges.

Considering the worldwide development community, SourceForge is also an important component of the infrastructure used by Chinese developers. More than 36,000 registered SourceForge users and 850 registered projects were estimated as Chinese.

Cosoft forge is, by far, the forge with more lines of code archived. However, the largest projects hosted in Cosoft are Linux distributions, which have been ignored in this study. Even with this exception, Cosoft is still the forge with more lines of code, more than 2,000,000. The next forge by lines of code is HitGforge, with about 400,000 lines of code.

Generally speaking, mailing lists are not the main channel of communication used in Chinese forges. In total, less than 150 mailing lists have been found. The OSDN and HitGForge forges, are the only ones that show more mailing lists than software releases or SCM (source code management) repositories.

Regarding SCM repositories, no more than 250 have been found (with just 82 if SourceForge is omitted). SCM repositories are spread in similar quantities among OSDN, HitGForge, Cosoft and OSS forges, despite the enormous differences in the number of registered users and registered projects.

#### 5.1.1 Forges analysed

The forges mentioned in the previous section, have been spidered in order to identify the projects they contain. Table 5.1 lists a relation of the number of projects and users registered in each of the forges. The

world's most popular forge, SourceForge, has been added to the table as many Chinese developers and Chinese-driven projects have been found there. The number of registered users at SourceForge gives the estimation of Chinese developers identified as such in SourceForge<sup>1</sup>. The 851 projects in SourceForge that are Chinese-driven have a majority (i.e. more than 50%) of Chinese developers in their teams.

Forge	Registered Users	Registered Projects
OSDN	3141	187
HitGforge	565	101
Cosoft	46834	1451
CNforge	no data	no data
OSS	6583	112
Sourceforge	36517	851

Table 5.1: Registered users and projects in Chinese forges (data 16th April 2007). SourceForge has been included for completeness (data June, 2006).

It is important to point out that not all registered users are active developers in the forges. Many of them could register and never join a development project, for instance. Projects, as well, may not make use of all development-related tools offered by the forges (and therefore, will not offer data in several kinds of repositories). This fact is shown in detail in the summary table 5.2, where the number of SCM (CVS/SVN) repositories, committers, commits, mailing lists, software releases and size of the software is given.

Forge	SCM repos	Committers	Commits	MailingLists	Releases	SLOC
OSDN	34/64	33	25,658	74	7/20	131,231
HitGforge	9/54	8	2,059	55	12/14	399,723
Cosoft	70/73	67	29,068	ND	48/59	2,090,299
CNforge	ND	ND	ND	ND	5/8	111,543
OSS	59/59	3	822	ND	8/13	127,478
SourceForge	82	82	24,053	ND	ND	ND

Table 5.2: Information sources that could be extracted from Chinese forges (April-May 2007).

Figure 5.1 shows the number of SCM repositories, mailing lists and software releases identified and analysed. Also, figure 5.2 and figure 5.3 show number of detected committers and commits respectively.

<sup>1</sup>Methodology report - Chapter Methodology - Section Global forge's analysis

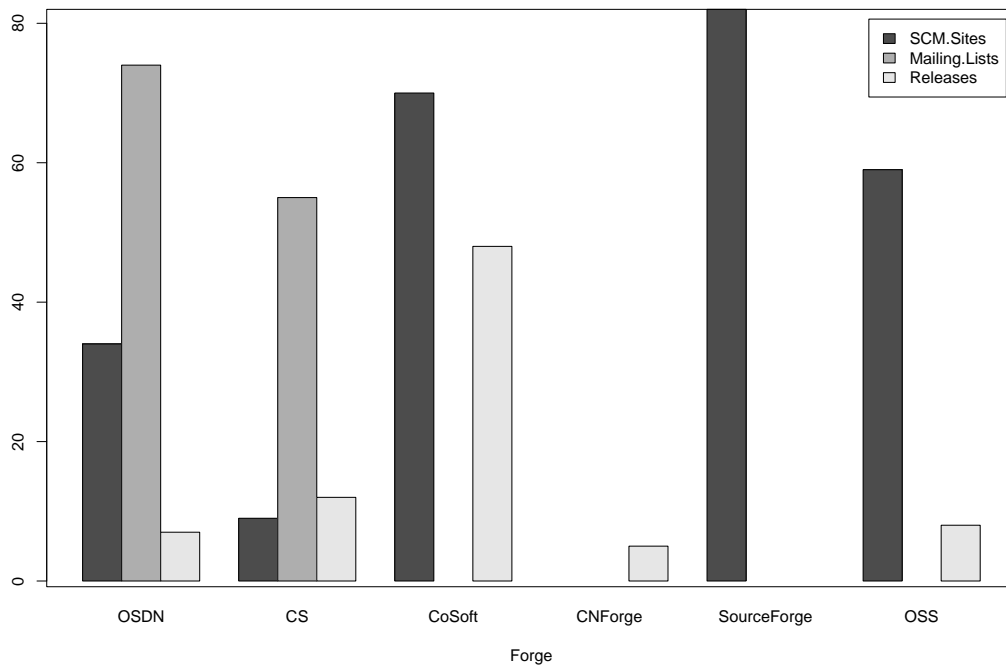


Figure 5.1: SCM repositories, mailing lists and software releases found in forges

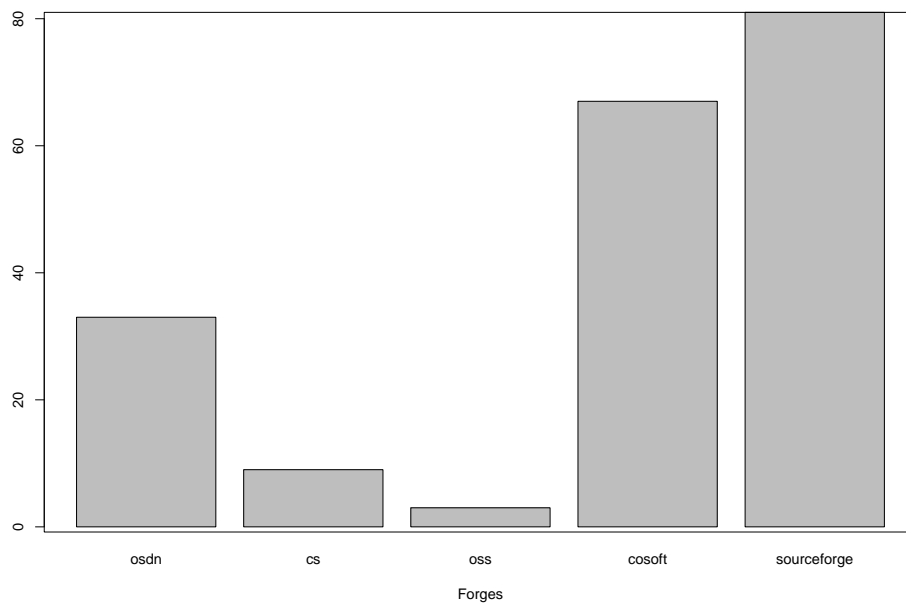


Figure 5.2: Committers per forge (estimated, in the case of SourceForge)

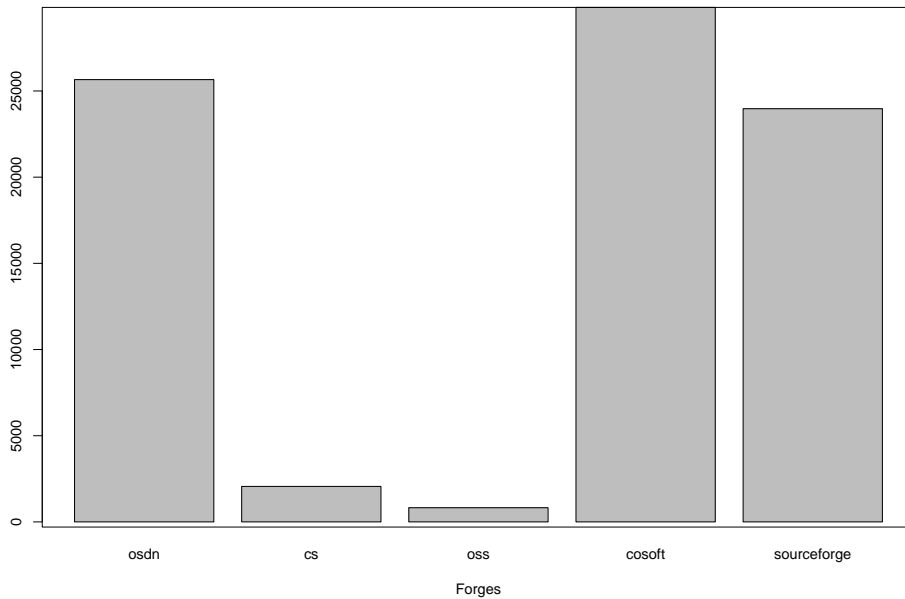


Figure 5.3: Commits per forge (estimated, in the case of SourceForge)

### 5.1.2 Programming languages

The programming languages used for writing the code found in the forges vary from forge to forge, although in general an overrepresentation of scripting languages (with respect to the distribution found in global software development projects) is found. As an example, table 5.3 and figure 5.4 contain information about the percentage of use of programming languages in Cosoft (not including results from the Linux distributions hosted in that forge).

Programming language	No. of lines	Percentage
C	718286	34.3628
C ++	326152	15.6031
PHP	304171	14.5515
Perl	261675	12.5185
Python	202526	9.6888
Pascal	170160	8.1404
Shell	52231	2.4987
Java	38942	1.8629
Asm	8475	0.4054
JSP	4689	0.2243
Lisp	2474	0.1183
Yacc	316	0.0151
Sed	202	0.0096
Total	2090299	100.0

Table 5.3: Programming languages used in Cosoft forge

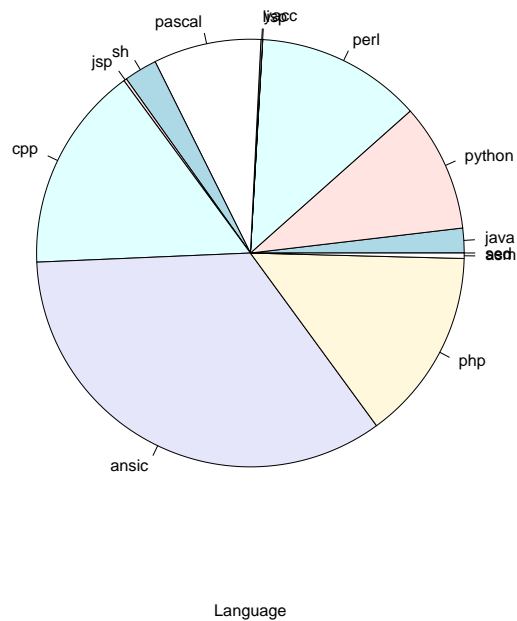


Figure 5.4: Programming languages used in Cosoft forge

### 5.1.3 Authorship data

The authorship data is based on the analysis of copyright notices in the source code. This analysis has been performed on the release distributions found in the forges, using some heuristics for grouping authors (or, to be more precise, copyright holders) in some large categories. Table 5.4 and figure 5.5 show these results.

Type of author	Detected number
Individual authors	267
Teams and Groups	22
Universities	17
Enterprises	90
Foundation or public entities	5
Anonymous/Unknown	11

Table 5.4: Type of authorship in Chinese forges

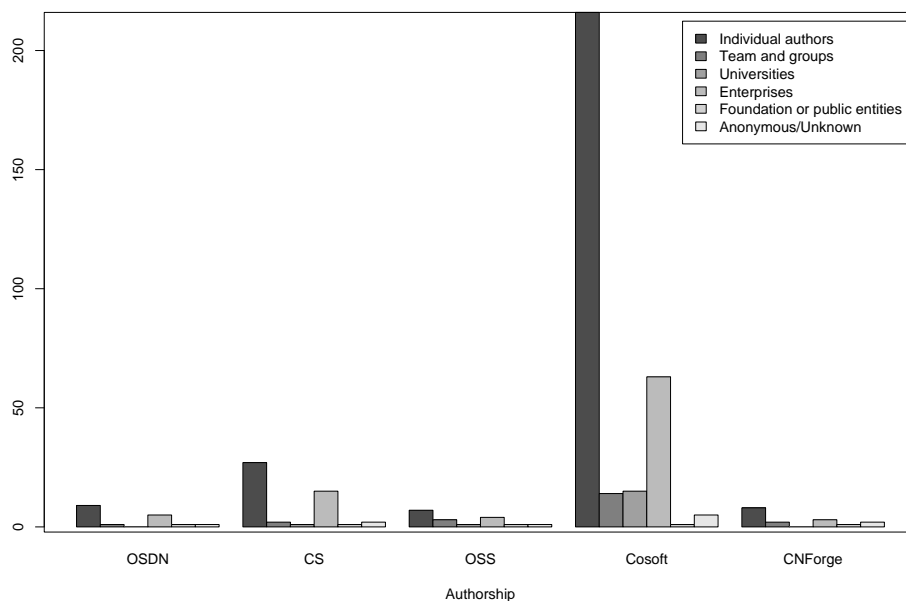


Figure 5.5: Authorship data (in software releases).

In this case, it is important to notice that despite the large number of users registered in forges, the number of authors is relatively small. In addition, some of them are not Chinese, since the code which is being modified or developed in the forges is in some cases, a derived version of some other software developed elsewhere.

#### 5.1.4 Other data

In addition to analysing forges, the study tried to find some other sources of information that could help in the identification of community-related parameters, and could also be a source for further analysis. As a result, the data shown in table 5.5 was collected.

Region	Communities	Lugs	Projects	Platforms
China	5	2	43	2

Table 5.5: Other data collected

In this table *Communities* means group of users interested in libre software, *LUGs* are Official Linux User Groups (with a physical address) and *Platforms* are web sites which provide any kind of support to the libre software world, such as forges.

## 5.2 India

Sarovar forge is the only local forge found in this study. It hosts around 450 projects and about 160 of them have analysable Source Code Management systems. The data shows that these projects have an average of two developers per project and that there is a total of 63 developers that have actually ever submitted code to the Source Code Management systems. Four of these projects have an important traffic of messages in their mailing lists. In fact, project *Devnag*<sup>2</sup> shows 36 different posters in its archives who

<sup>2</sup><http://devnag.sarovar.org/>

sum up to more than 600 messages in total.

Some other general data can be obtained crossing numbers. For instance, around 2% of registered people in Sarovar make commits. All in all, in Sourceforge, around three percent of registered users submit code to the Source Code Management systems.

While SourceForge is a global forge which hosts projects of all the world, Sarovar is the local Indian forge with the most registered users. This forge is not as big as SourceForge but it has 461 registered users and 3237 projects, while SourceForge has around 1383 users and 22113 projects that have been identified as Indian. But as stated before, just a small part of the total number of users really submit code to the Source Code Management systems (in this case CVS). Precisely, 63 committers in Sarovar and 138 Indian committers in SourceForge.

Precise numbers about the CVS activity show that although SourceForge has more than 3 times more registered projects than Sarovar, the truth is that Sarovar's developers have made more than the double of commits than SourceForge's Indian developers. Around 72,355 in Sarovar and 28,749 in Sourceforge. This is because the number of projects that use the Source Code Management tools in each forge is very similar.

Regarding to Linux User Groups (LUGs) and communities, around two hundred and thirty of them have been found (data provided by partners). The widest spread language among them is English and their favourite platform is Yahoo groups.

Also, some companies have been found participating in libre software projects. A classification of the authorship of the commits in the forges, show that there are fifty five different Indian authors (from partners data) and they are divided in five main groups. Forty two have been classified as Individual authors, three of them work in team or group, four of them work in an University, three of them come from enterprises and one has been identified as a foundation or a public entity.

### 5.2.1 Forges analysed

The only Indian local forge identified (Sarovar), has been spidered in order to identify the projects it contains. Table 5.6 shows registered users and projects in Indian forges (Data 16th April 2007). SourceForge has been included for completeness (Data June 2006). project in

Forge	Registered Users	Registered Projects
Sourceforge	22113	1383
Sarovar	3237	461

Table 5.6: Registered users and projects in Indian forges, including SourceForge

It is important to point out that not all registered users are active developers in the forges. Many of them could register and never join a development project, for instance. Projects, as well, may not make use of all development-related tools offered by the forges (and therefore, some kinds of data sources, may offer no data at all). This fact is clearly revealed in the summary table 5.7. Table 5.7 shows the number of SCM (CVS/SVN) repositories, commits, mailing lists, messages, posters, software releases and size of the software.

Forge	CVS/SVN Sites	Commits	MailingLists	Messages	Posters	Releases	SLOC
Sarovar	153/167	72,355	18	925	102	70/91	74,037
SourceForge	130	28,749	ND	ND	ND	ND	ND

Table 5.7: Information sources that could be studied in Indian forges (April 2007).

Figure 5.6 shows the number of SCM repositories, mailing lists and software releases identified and analysed. Also, figure 5.7 and figure 5.8 show the number of detected committers and commits in some

selected projects respectively.

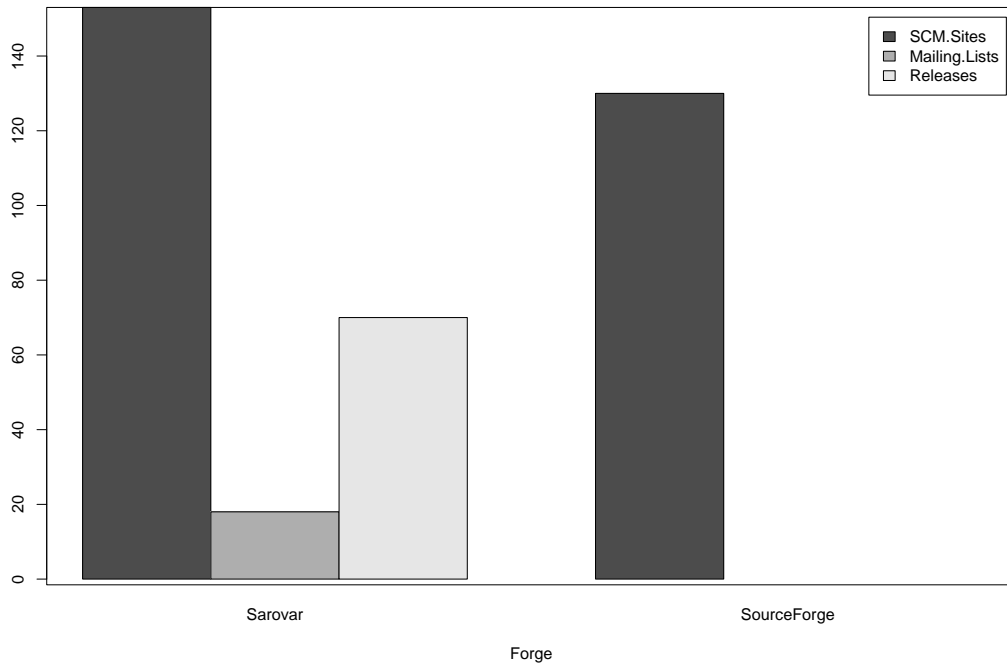


Figure 5.6: SCM repositories, mailing lists and software releases found in forges

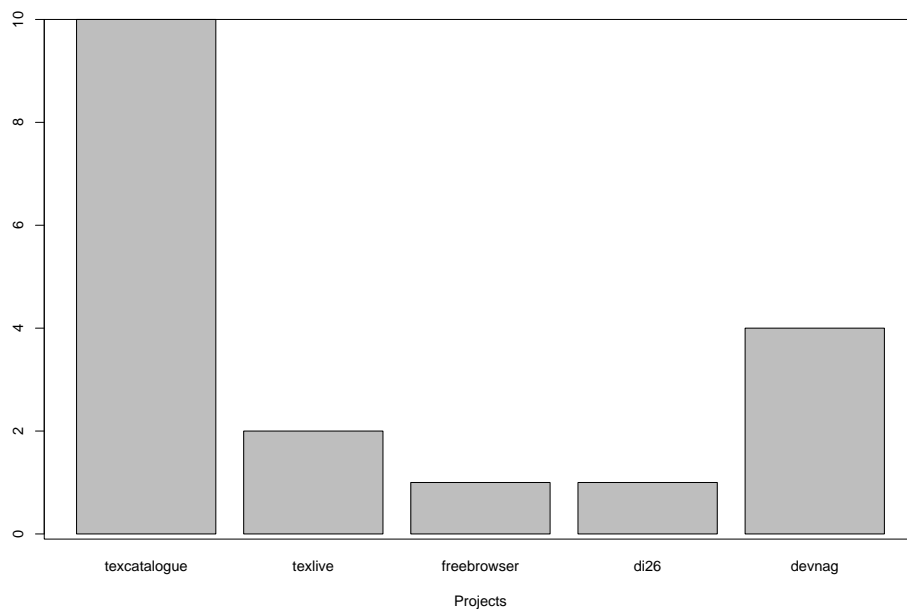


Figure 5.7: Committers in some selected projects

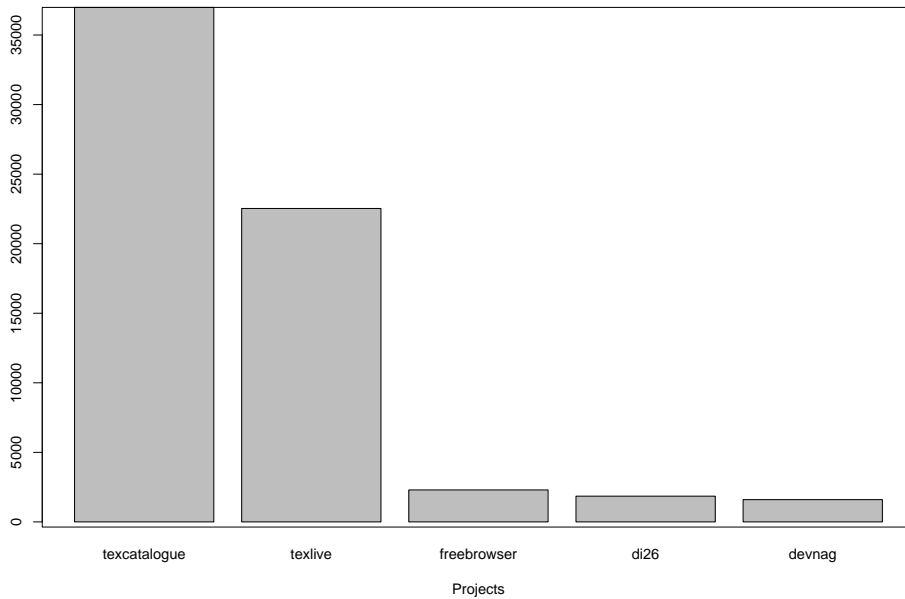


Figure 5.8: Commits in some selected projects

## 5.2.2 Programming languages

The programming languages used for writing the code found in forges vary from forge to forge and from project to project. Table 5.8 and figure 5.9 contain information about the percentage of use of programming languages in Sarovar.

Programming Language	N. of Detected Lines	Percentage
C	137849	40.4197
C ++	87104	25.5403
Sh	44843	13.1487
Python	25515	7.4814
PHP	18531	5.4336
Perl	8343	2.4463
Lisp	4190	1.2285
Tcl	3533	1.0359
Java	3509	1.0288
Yacc	2466	0.7230
Pascal	1716	0.5031
Ruby	1306	0.3829
Asm	1085	0.3181
Lex	642	0.1882
Awk	235	0.0689
Sed	172	0.0504
C #	5	0.0014
Total	341044	100.0

Table 5.8: Programming Languages used in Sarovar forge



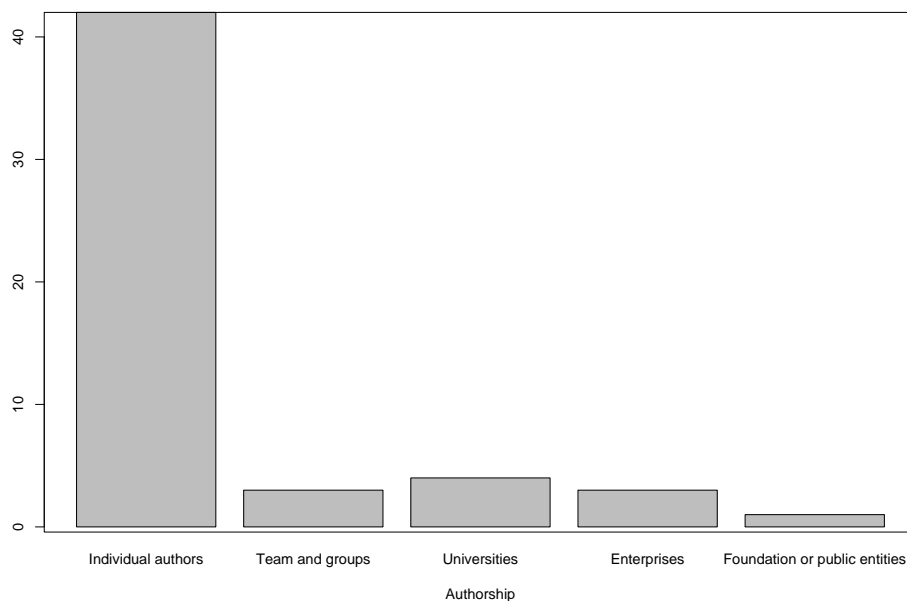


Figure 5.10: Authorship data (in software releases).

## 5.2.4 Other data

In addition to analysing forges, the study tried to find some other sources of information that could help in the identification of community-related parameters, and could also be a source for further analysis. As a result, the data shown in table 5.10 was collected.

Region	Communities	Developers	Lugs	Media	Platforms	Projects
India	4	171	233	2	2	79

Table 5.10: Other data collected

In this table *Communities* means group of users interested in libre software, *LUGs* are Official Linux User Groups (with a physical address), *Media* are any kind of journal magazine related to Linux and *Platforms* are web sites which provide any kind of support to the libre software world, such as forges.

## 5.3 South Africa

Two local forges (Avoir and Knowledge Tree) have been found in South Africa. These two forges host the development of libre software projects in this country. There is also a community of people identified as South Africans who use SourceForge for the same purpose, and this fact has been taken into account in this study.

Regarding to the two local forges, Knowledge Tree appears to be the one with the most registered users, around 310. On the other hand, we could not find out how many registered users Avoir has. However Avoir is the forge with the most registered projects. Avoir has 70 registered projects and Knowledge Tree has 51. The data from South African activity in SourceForge shows around 5700 South African users registered there and nearly 500 projects which were identified as South African-driven.

The data gathered during Source Code Management systems mining, show that Avoir has 22 repositories and 71 different committers using them. In this case, Knowledge Tree is a black box because

automatic data retrieval is not possible in their Source Code Management systems. Moreover, just 73 of the 494 South African projects in SourceForge use the Source Code Management system tool provided by the platform. Despite the difference in the number of repositories mined in Avoir (22) and in SourceForge (73), there is much more source code activity in Avoir than in SourceForge's South African community. Avoir has a history of around 92,888 commits while SourceForge shows 53,142.

It is also remarkable that the most popular programming language in the South African local forges is, by far, php. Php represents 82% of the source code in Avoir (followed by python) and 98% in Knowledge Tree (followed by C#). The total lines of code the two forges sum is 244,222.

### 5.3.1 Forges analysed

As stated before, two forges (Avoir and Knowledge Tree) should be considered as South African local forges. These forges have been spidered in order to identify the projects they contain. Table 5.11 lists a relation of the number of projects and users registered in each of the forges. The world's most popular forge, SourceForge, has been added to the table as many South African developers and South African-driven projects have been found there. The number of registered users in SourceForge should be understood as the estimation of South African developers<sup>3</sup> in SourceForge. The 494 projects in SourceForge that are identified as South African-driven have a majority (i.e. more than 50%) of South African developers in their teams.

Forge	Registered Users	Registered Projects
Avoir	no data	70
Knowledge Tree	308	51
Sourceforge	5706	494

Table 5.11: South African forges (on 16th of April of 2007)

It is important to point out that not all registered users are active developers in the forges. Many of them could register and never join a development project, for instance. Projects, as well, may not make use of all development-related tools offered by the forges (and therefore, some kinds of data sources, may offer no data at all). This fact is clearly revealed in the summary table 5.12. This table ( 5.12) shows the number of SCM (CVS/SVN) repositories, committers, commits, mailing lists, software releases and size of the software.

Forge	SCM Repos.	Committers	Commits	MailingLists	Releases	SLOC
Avoir	22	71	92,888	14	8	203286
Knowledge Tree	ND	ND	ND	ND	39	40936
Sourceforge	73	102	53,142	ND	ND	ND

Table 5.12: South African forges (on April-June of 2007). SourceForge has been included for completeness (data June, 2006)

Figure 5.11 shows the number of SCM repositories, mailing lists and software releases identified and analysed. Also, figure 5.12 and figure 5.13 show the number of detected committers and commits in forges.

<sup>3</sup>Methodology report: How to obtain nationality from Sourceforge

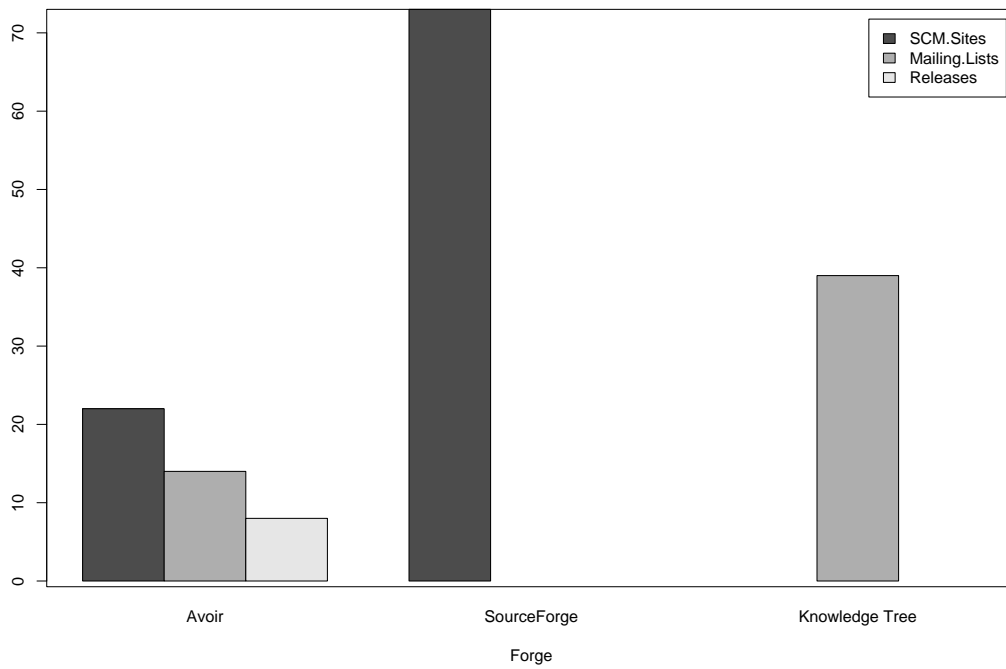


Figure 5.11: SCM repositories, mailing lists and software releases found in forges

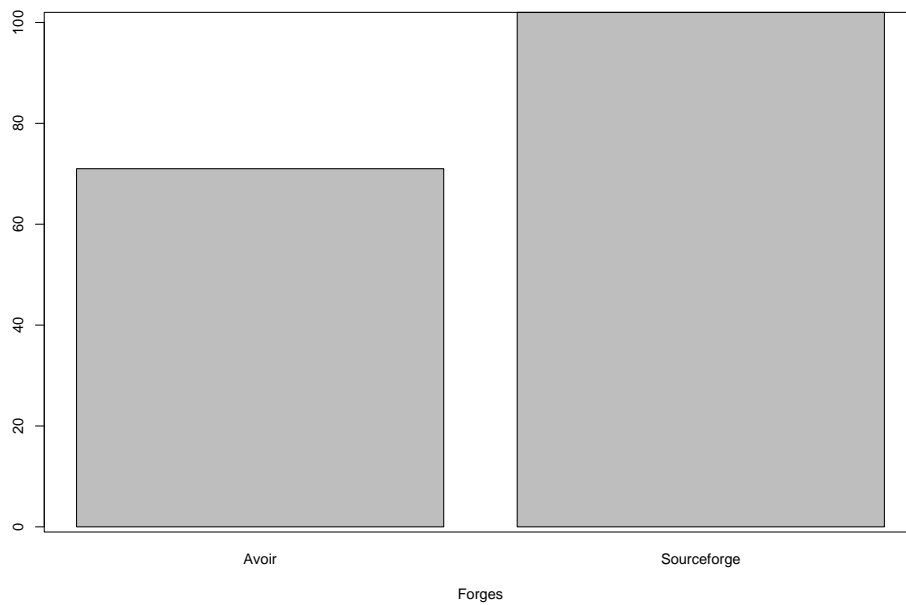


Figure 5.12: Committers per forge

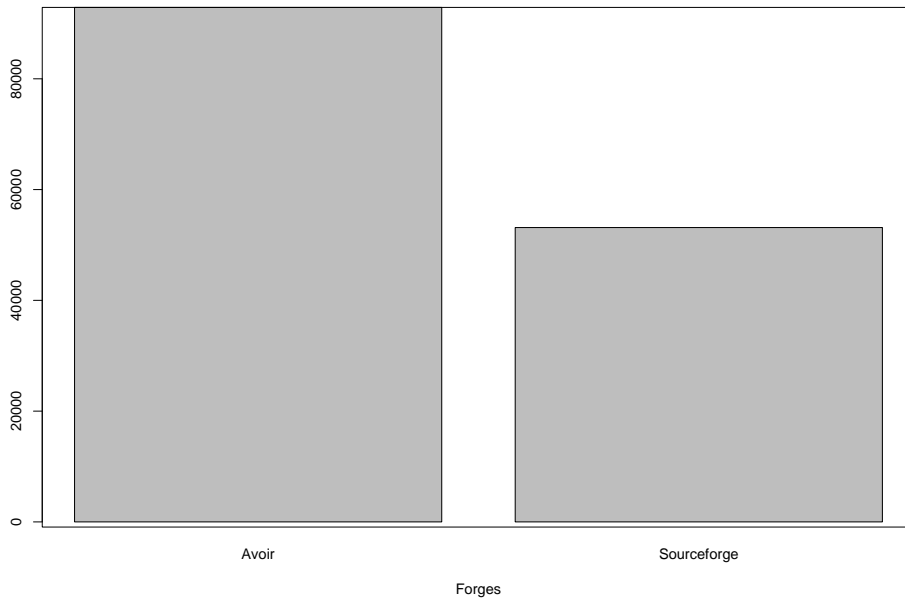


Figure 5.13: Commits per forge

### 5.3.2 Programming Languages

The programming languages used for writing the code found in forges vary from forge to forge and from project to project. Table 5.13 and figure 5.14 contain information about the percentage of use of programming languages in Avoir.

Programming language	No. of lines	Percentage
PHP	167512	82.4021
Python	12837	6.3147
Java	11964	5.8853
C	3113	1.5313
C #	3101	1.5254
JSP	2092	1.0290
SH	1337	0.6576
Perl	862	0.4240
Pascal	44	0.0216
C ++	424	0.2085
Total	203286	100.0

Table 5.13: Programming languages used in Avoir forge

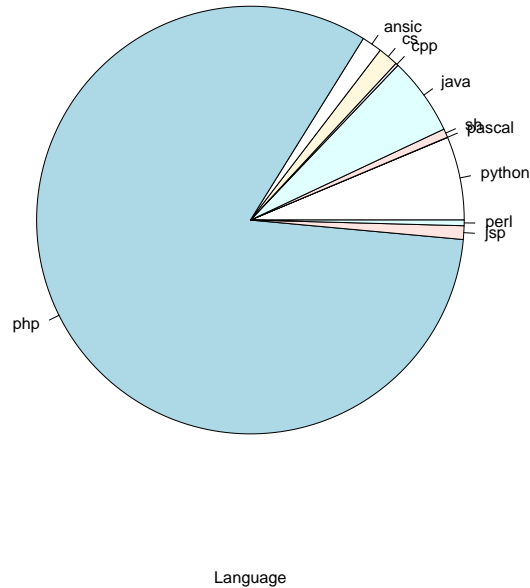


Figure 5.14: Programming languages used in Avoir forge

### 5.3.3 Authorship

The authorsip data is based on the analysis of copyright notices in the source code. This analysis has been performed on the release distributions found in the forges, using some heuristics for grouping authors (or, to be more precise, copyright holders) in some large cathegories. Table 5.14 and figure 5.15 show these results.

Type of author	Detected number
Individual author	52
Teams and groups	2
Universities	4
Enterprises	9
Foundation or public entities	1
Anonymous/Unknown	3

Table 5.14: Type of authorship in South African forges

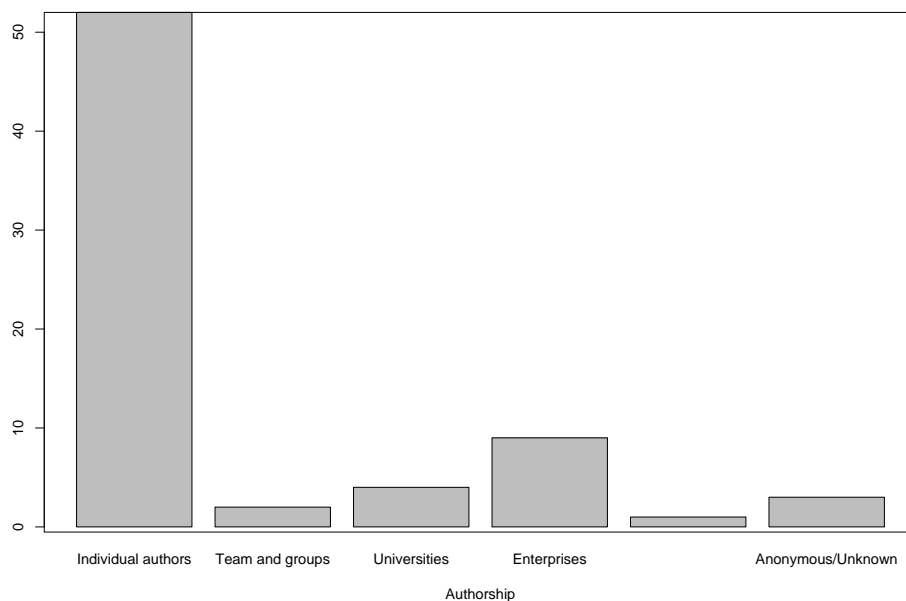


Figure 5.15: Authorship data (in software releases).

### 5.3.4 Other data

In addition to analysing forges, the study tried to find some other sources of information that could help in the identification of community-related parameters, and could also be a source for further analysis. As a result, the data shown in table 5.15 was collected.

Region	Communities	Lugs	Platforms	Projects
South Africa	3	3	14	21

Table 5.15: Other data collected

In this table *Communities* means group of users interested in libre software, *LUGs* are Official Linux User Groups (with a physical address) and *Platforms* are web sites which provide any kind of support to the libre software world, such as forges.

## 5.4 Brazil

Five local forges, which are quite popular among the libre software community, have been found. From these 5 forges, the most important one is *CodigoLivre* which has around 14200 registered users and 1866 projects. Then comes *LuaForge* with around 800 users and 250 projects and *AgroLivre* with 615 users and just 16 projects. On the other hand, we do not have enough information about the last two of them, *Incubadora Virtual* and *QuerenciaLivre*. Also Brazilian developers use SourceForge to collaborate in libre software projects and to develop Brazilian projects too. We identified 851 Brazilian projects in SourceForge and it is remarkable that *CodigoLivre* has even more.

In *CodigoLivre* there are 96 mailing list including general mailing list and specific projects mailing lists, with a total of around 1667 different posters. We can say that this mailing lists are very active because they had a total of more than 5700 messages posted when the data was retrieved, in January 2007.

It is worth to mention that there are at least 9 enterprises currently submitting code to the forges in Brazil. This shows a certain interest from Brazilian industries in SourceForge.

Comparing the Source Code Management systems of the local forges and SourceForge we see that projects hosted in local forges have a valid SCM tool but only around 640 of the 1840 use it to develop. Only 643 developers use their SCM account regularly. Also, there are 300 projects in SourceForge with a CVS repository in use. About 414 developers are working on these repositories.

### 5.4.1 Forges analysed

The Brazilian forges have been spidered in order to identify the projects they contain. Table 5.16 lists a relation of the number of projects and users registered in each of the forges. The world-wide most popular forge, SourceForge, has been added to the table as many Brazilian developers and Brazilian-driven projects have been found there. The number of registered users at SourceForge gives only the estimation of Brazilian developers<sup>4</sup>. The 2560 projects in SourceForge that are Brazilian-driven have a majority (i.e. more than 50%) of Brazilian developers in their teams. Figure 5.16 shows data about projects and users per forge.

Forge	Registered users	Registered Projects
CodigoLivre	14200	1866
LuaForge	802	250
AgroLivre	615	16
IncubadoraVirtual	ND	ND
QuerenciaLivre	ND	ND
Sourceforge	21291	851

Table 5.16: Registered users and projects in Brazilian forges (30th of May 2007). SourceForge has been included for completeness (data June, 2006).

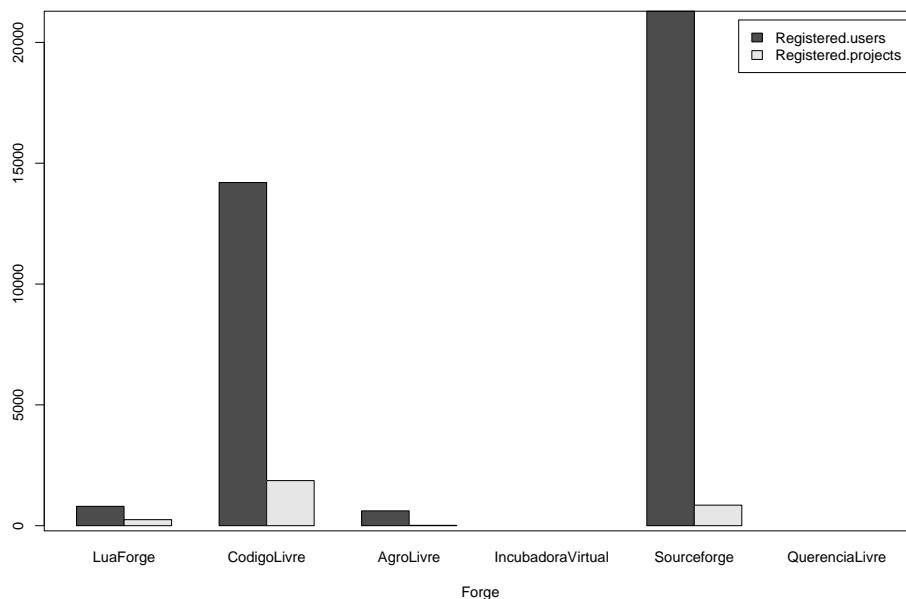


Figure 5.16: Projects and users per forge

<sup>4</sup>Methodology report: How to obtain nationality from Sourceforge

It is important to point out that not all registered users are active developers in the forges. Many of them could register and never join a development project, for instance. Projects, as well, may not make use of all development-related tools offered by the forges (and therefore, will not offer data from several data sources). This fact is shown in detail in the summary table 5.17, where the number of SCM (CVS/SVN) repositories, mailing lists, software releases and size of the software is given.

This fact is specially dramatic for Brazil. Being one of the countries of the study with the most local forges (5), it is specially disappointing that not much data could be extracted from them.

Forge	SCM repos	MailingLists	Releases	SLOC
CodigoLivre	1950	96	89	5,330,730
LuaForge	ND	ND	ND	ND
AgroLivre	ND	ND	ND	ND
IVirtual	ND	ND	ND	ND
QLivre	ND	ND	ND	ND
SourceForge	300	414	ND	ND

Table 5.17: Information sources that could be extracted from Brazilian forges (April 2007).

## 5.4.2 Programming Languages

The programming languages used for writing the code found in forges vary from forge to forge and from project to project. Table 5.4.2 and figure 5.17 contain information about the percentage of use of programming languages in *CodigoLivre*.

Programming Language	Lines of Code	Percentage
C	4390322	82.3587
PHP	365592	6.8581
ASM	248461	4.6609
SH	63580	1.1927
C++	61519	1.1540
F90	58361	1.0948
Python	57034	1.0699
Pascal	44525	0.8352
Perl	22825	0.4281
Java	9851	0.1847
Yacc	6219	0.1166
Lex	1513	0.0283
Ruby	275	0.0051
Lisp	218	0.0040
C#	211	0.0039
Awk	99	0.0018
Sed	93	0.0017
TCL	32	0.0006
Total	5330730	100.0

Table 5.18: Programming Languages used in CodigoLivre forge

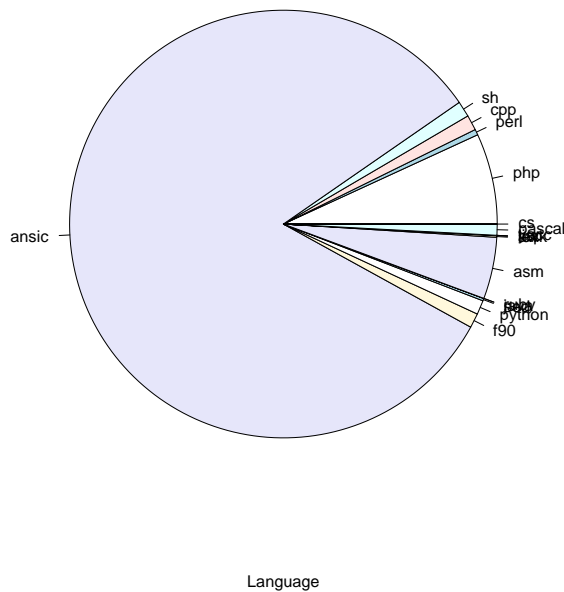


Figure 5.17: Programming languages used in *CodigoLivre* forge

### 5.4.3 Authorship

The authorsip data is based on the analysis of copyright notices in the source code. This analysis has been performed on the release distributions found in the forges, using some heuristics for grouping authors (or, to be more precise, copyright holders) in some large categories. Table 5.19 and figure 5.18 show these results.

Type of author	Detected number
Individual authors	26
Team-Groups	5
Universities	1
Enterprises	9
Foundation or public entities	1
Unknown	4

Table 5.19: Type of authorship in Brazilian forges

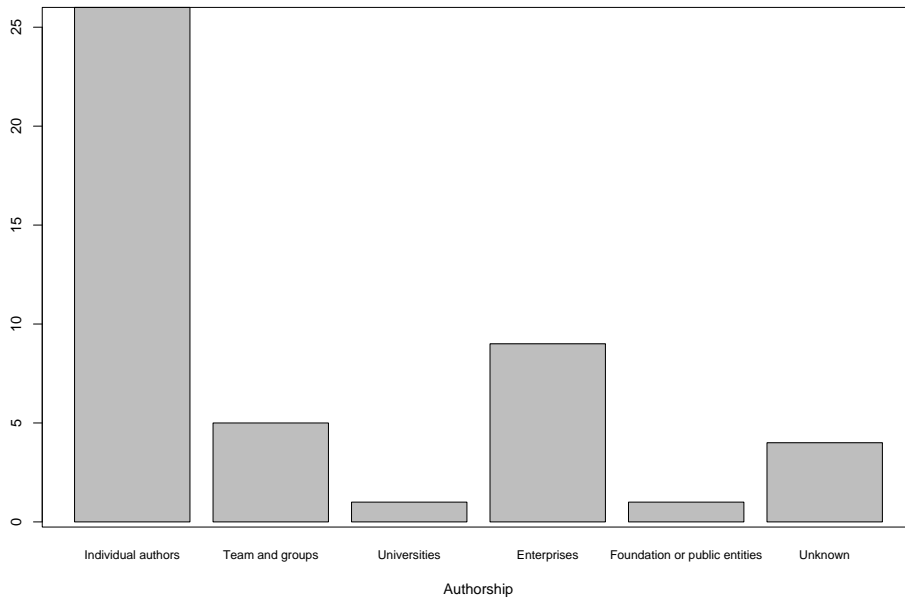


Figure 5.18: Authorship data (in software releases).

#### 5.4.4 Other data

In addition to analysing forges, the study tried to find some other sources of information that could help in the identification of community-related parameters, and could also be a source for further analysis. As a result, the data shown in table 5.20 was collected.

Region	Communities	Media	Platforms
Brazil	20	1	1

Table 5.20: Other data collected

In this table *Communities* means group of users interested in libre software, *Media* are any kind of journal magazine related to Linux and *Platforms* are web sites which provide any kind of support to the libre software world, such as forges.

### 5.5 Argentina

Besides SourceForge (the largest forge in the world, for which the Argentinian participation has been estimated), this study has analysed one Argentinian forge. LugFi is the only local forge in Argentina that could be identified. It hosts more than 155 registered users and more than 44 registered projects.

Considering the worldwide development community, SourceForge is the most important component of the infrastructure used by Argentinian developers. More than 5439 registered SourceForge users and 849 registered projects were identified as Argentinian.

In the LugFi forge there are 15 projects from which we have gathered relevant data. The projects that have a larger number of source code lines are *arrakis*, *dapplet* and *powertrans*. It is possible to find 8 different programming languages in LugFi Forge, being “ANSI C” the most popular.

Generally speaking, mailing lists are not the main channel of communication used by Argentinian developers. In total, there are 6 mailing lists and all of them found in the LugFi forge.

Looking at the use of Source Code Management systems, some remarkable data shows up. Only a few of the projects, which the forge is hosting, use the Source Code Management tools provided, in this case CVS. From the 44 projects just 34 of them have created a CVS tree, and only 14 use this CVS for development regularly.

The study of the authorship of the source code of the projects shows that most of the projects appear to be developed by individual authors. We did not find evident trace of code from enterprises, universities or organisations.

### 5.5.1 Forges analysed

The LugFi forge has been spidered in order to identify the projects it contains. Table 5.21 lists a relation of the number of projects and users registered in each of the forges. The world's most popular forge, SourceForge, has been added to the table as many Argentinian developers and Argentinian-driven projects have been found there. The number of registered users at SourceForge is actually the estimation of Argentinian developers identified as such in SourceForge<sup>5</sup>. The 849 projects in SourceForge that are Argentinian-driven have a majority (i.e. more than 50%) of Argentinian developers in their teams.

Forge	Registered Users	Projects
SourceForge	5439	849
LugFi	155	44

Table 5.21: Registered users and projects in Argentinian forges (April 2007). SourceForge has been included for completeness (data June, 2006).

It is important to point out that not all registered users are active developers in the forges. Many of them could register and never join a development project, for instance. Projects, as well, may not make use of all development-related tools offered by the forges (and therefore, will not offer data in several kinds of repositories). This fact is shown in detail in the summary table 5.22, where the number of SCM (CVS/SVN) repositories, committers, commits, mailing lists, software releases and size of the software is given.

Forge	SCM repos	Committers	Commits	MailingLists	Releases	SLOC
LugFi	16/36	16	4,620	6	15/17	63,349
SourceForge	107	132	38,481	ND	ND	ND

Table 5.22: Information sources that could be extracted from Argentinian forges (April-May 2007).

Figure 5.19 shows the number of SCM repositories, mailing lists and software releases identified and analysed. Also, figure 5.20 and figure 5.21 show number of detected committers and commits respectively.

---

<sup>5</sup>Methodology report - Chapter Methodology - Section Global forge's analysis

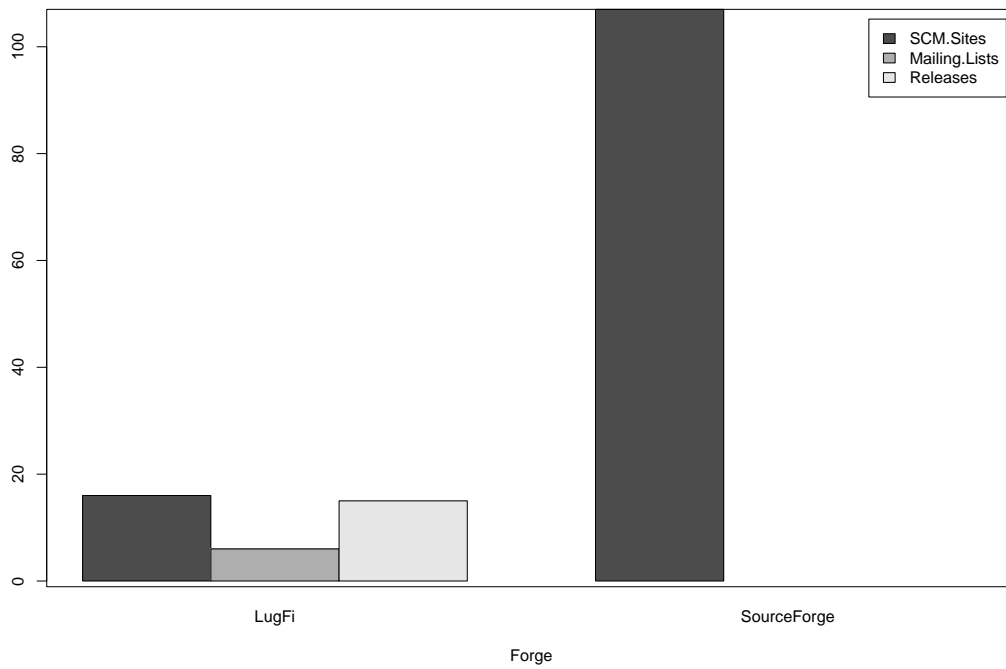


Figure 5.19: SCM repositories, mailing lists and software releases found in forges

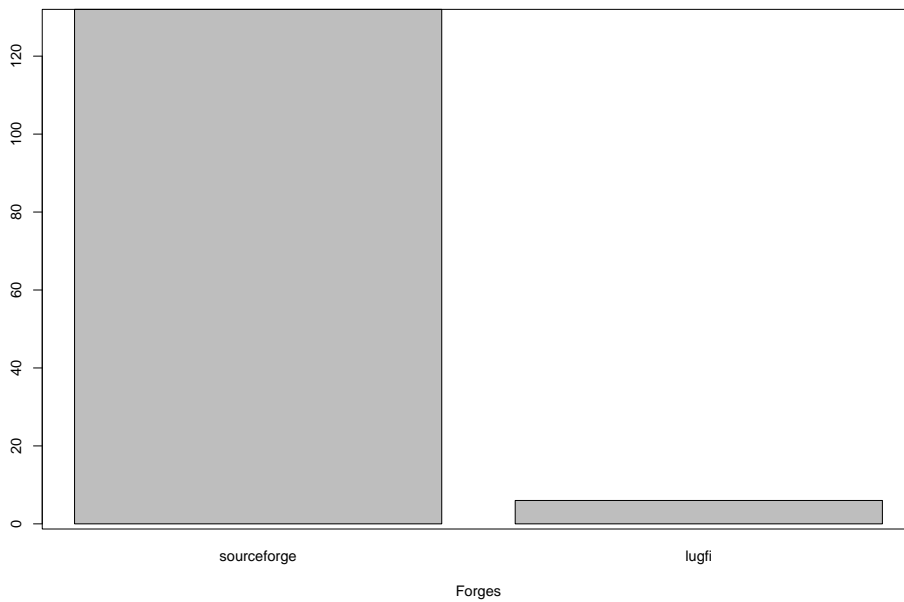


Figure 5.20: Committers per forge

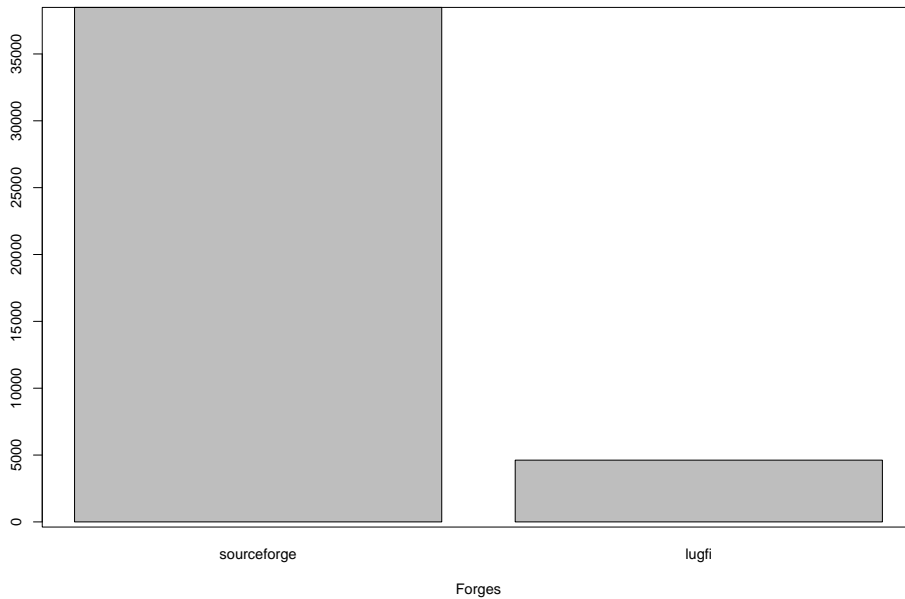


Figure 5.21: Commits per forge

## 5.5.2 Programming Languages

The programming languages used for writing the code found in forges vary from forge to forge and from project to project. Table 5.23 and figure 5.22 contain information about the percentage of use of programming languages in *LugFi*.

Programming language	N. of detected lines	Percentage
ansic	25957	43.5389
c++	13804	23.1541
sh	10245	17.1844
php	8382	14.0600
perl	419	0.7028
awk	386	0.6474
python	332	0.5569
sed	93	0.1560

Table 5.23: General Language results

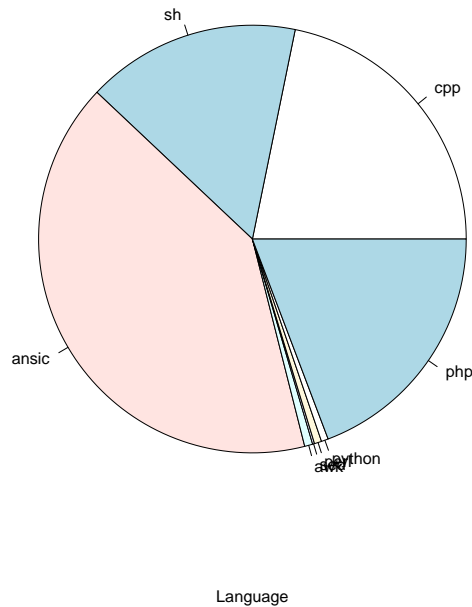


Figure 5.22: Information of languages used in the develop of projects

### 5.5.3 Authorship

The authorsip data is based on the analysis of copyright notices in the source code. This analysis has been performed on the release distributions found in the Argentinian forges, using some heuristics for grouping authors (or, to be more precise, copyright holders) in some large cathegories. Table 5.24 shows these results.

Type of author	Detected number
Individual authors	327
Unknown	49

Table 5.24: Type of authorship in Argentinian forges

### 5.5.4 Other data

In addition to analysing forges, the study tried to find some other sources of information that could help in the identification of community-related parameters, and could also be a source for further analysis. As a result, the data shown in table 5.25 was collected.

Region	Communities	Developers	Lugs	Projects	Platforms
Argentina	7	126	57	2	1

Table 5.25: Other data collected

In this table *Communities* means group of users interested in libre software, *LUGs* are Oficial Linux User Groups (with a physical address) and *Platforms* are web sites which provide any kind of support to the libre software world, such as forges.

## 5.6 Malaysia

There is no Malaysian local forge equivalent to SourceForge but the Open Source Competency Centre (OSCC) Knowledge Bank, which is owned and managed by Malaysian Administrative Modernisation and Management Planning Unit (MAMPU), Prime Minister's Department, facilitates OSS development activities as part of the Malaysian Public Sector OSS Master Plan. OSCC has around 2016 registered users and 26 active projects initiated mainly by Public Sector agencies. The main focus of OSCC Knowledge Bank is to enable the sharing of source codes, knowledge and experience to facilitate and support greater adoption and development of OSS within the Public Sector agencies. In addition, OSCC also serves as a one stop reference centre that bridges the private sector, R&D and public sector communities together in an effort to match the supply and demand of OSS. To this effect, there is a repository of 250 OSS solutions and 201 OSS service providers. Malaysian developers also use Sourceforge to develop their projects. There are around 3189 registered users in SourceForge that the team identified as Malaysians and 94 projects that appear to be Malaysian-driven. However, not all registered users are developers.

Regarding the mailing list, there are a considerable number ( 12 ) LUGs in Malaysia, and there are 11 Media mainly from computer magazines and also from main stream newspapers with weekly computer columns covering OSS news and articles in the country. Majority are written in English, but there are Bahasa Malaysia and also Chinese languages articles from the press

Looking at the authorship of source code, most of the projects are developed by individuals. Followed by enterprises, who sustain the second biggest group.

In the *OSCC* Knowledge Bank, Source Code Management Tools are not applicable. On the other hand, SourceForge developers use the version control system (CVS) provided by this platform to develop the software for the project. We can find 108 projects with this active CVS repositories and they all sum up to 60,365 commits done in SourceForge.

There are 3 general mailing lists in *OSCC* but they do not have much movement with only 17 different email addresses within them.

Regarding to source code information, Malaysian developers seem to prefer to develop in *shell script* and *php* more than in other programming languages. These two languages represent 87% of the total lines of source code.

### 5.6.1 Forges analysed

The Malaysian FLOSSWorld partners have identified one Malaysian OSS Repository named *OSCC Knowledge Bank*. For technical reasons, the spider analysis did not work as expected in OSCC. But manual inspection showed that it hosts 26 projects.

Table 5.26 lists a relation of the number of projects and users registered in the forges. The world's most popular forge, SourceForge, has been added to the table as many Malaysian developers and Malaysian-driven projects have been found there. The number of registered users at SourceForge gives the estimation of Malaysian developers identified as such in SourceForge<sup>6</sup>. The 94 projects in SourceForge that are Malaysian-driven have a majority (i.e. more than 50%) of Malaysian developers in their teams.

Forge	Registered Users	Projects
Sourceforge	3189	94
Osc	2016	26

Table 5.26: Registered users and projects in Malaysian forges (data 16th April-May 2007). SourceForge has been included for completeness (data June, 2006).

<sup>6</sup>Methodology report - Chapter Methodology - Section Global forge's analysis

It is important to point out that not all registered users are active developers in Sourceforge and OSCC Knowledge Bank. For example, many of them could register and not join a development project. Since OSCC Knowledge Bank was established for the purpose of sharing information, source code and software programs among agencies and does not make use of all the tools for developers, it cannot be taken as an apple-to-apple comparison with Sourceforge. Therefore, the relevant data for SCM, Committers and Commits for projects in OSCC Knowledge Bank will not be available currently. In Phase II of the Malaysian Public Sector OSS Master Plan, OSCC Knowledge Bank will provide more tools to facilitate and support increased development activities.

Forge	SCM repos	Committers	Commits	MailingLists	Releases	SLOC
Osc	NA	NA	NA	3	21	116,065
SourceForge	108	183	60,365	ND	ND	ND

Table 5.27: Information sources that could be extracted from Malaysian forges (April-May 2007).

## 5.6.2 Programming Languages

The programming languages used for writing the code found in forges vary from forge to forge and from project to project. Table 5.23 and figure 5.22 contain information about the percentage of use of programming languages in *OSCC Knowledge Bank*.

Programming language	No. of lines	Percentage
SH	61,872	53.308%
PHP	40,298	34.7201%
Perl	5,585	4.8119%
Java	4,002	3.448%
C	3,820	3.2912%
Awk	170	0.1464%
Python	167	0.1438%
C++	118	0.1016%
Pascal	33	0.0284%
Total	116,065	100%

Table 5.28: Programming languages used in *Osc Knowledge Bank*

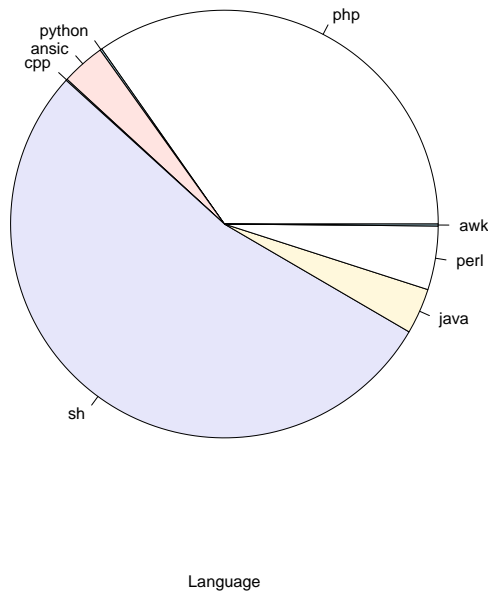


Figure 5.23: Programming languages used in *Osc* Knowledge Bank

### 5.6.3 Authorship

There are different types of authors in each project. Some of them, work for companies that are interested in participating in the project. Some of them are individual volunteers. Others are R&D Institutes or Universities. The results in this section have been obtained from software releases in *OSCC* (not from SCM repositories), by analysing copyright attributions in source files. Different authors identified have been classified into categories. Table 5.29 and figure 5.24 show these results.

Type of author	Detected number
Individual authors	86
Team-Groups	6
Universities	3
Enterprises	9
Foundation or public entities	1
Unknown	3

Table 5.29: Type of authorship in Malaysian forges

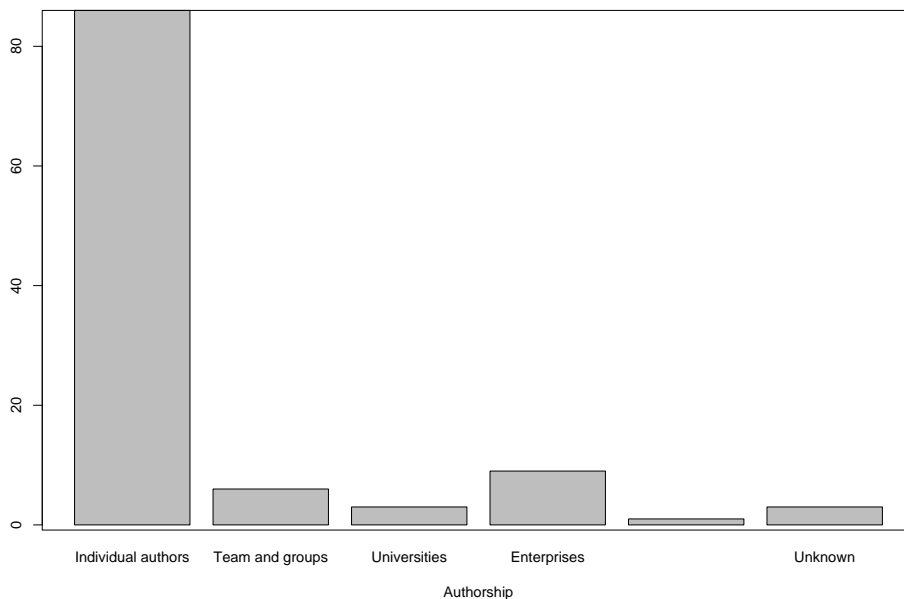


Figure 5.24: Authorship data (in software releases).

#### 5.6.4 Other data

In addition to analysing forges, the study tried to find some other sources of information that could help in the identification of community-related parameters, and could also be a source for further analysis. As a result, the data shown in table 5.30 was collected.

Region	Communities	Developers	Lugs	Media	Projects	Platforms
Malaysia	17	123	12	11	41	17

Table 5.30: Other data collected

In this table *Communities* means group of users interested in libre software, *LUGs* are Oficial Linux User Groups (with a physical address), *Media* are any kind of journal magazine related to Linux and *Platforms* are web sites which provide any kind of support to the libre software world, such as forges.

### 5.7 Croatia

The development and wide-spread implementation of Free/Libre Open Source Software in Croatia is not very prominent yet. Its introduction is in a early stage, and currently, there is not a big libre software community in the country. Therefore, the results in this report are not as complete as in other studies performed during the FLOSSWorld project, but still represent a valuable information.

This study found one Croatian local forge named *Linux.hr*, from which it has not been possible to obtain neither the total number of registered users nor the total number of hosted projects. Despite of this drawbacks, 7 projects out of a total of 11 projects found, fulfilled the requirements needed to perform the analysis. There are 6 different committers in *Linux.hr*, and all of them cooperate in more than one project. We cannot tell if they are real people or organizations. However, all of them together generated 12400 lines of code that are hosted in the forge.

Looking at the worldwide development community, it seems that a great number of croatian users are reunited around SourceForge. More than a 1200 users and 104 projects were identified as Croatians in SourceForge.

The *Linux.hr* forge stores 14 mailing lists and it is noticeable that just 2% of the contributors generate most of the traffic of the list. Just 24 people out of 1200 have ever written more than 10 messages.

A total of 19 Croatian-driven Source Code Management systems were found (12 from SourceForge and 7 from *Linux.hr*). It is remarkable that almost 90% of the croatian projects hosted in SourceForge do not have a Source Code Management system.

### 5.7.1 Forges analysed

*Linux.hr* is the website of the Croatian Linux Association, which is the national Linux Users Group: HULK, and it hosts a forge. This forge has been spidered in order to identify the projects it contains. However, we could not obtain an estimation of the number of registered users and projects in it. The world's most popular forge, SourceForge, has been added to the table as many Croatian developers and Croatian-driven projects have been found there. In Table 5.31 the number of registered users at SourceForge should be understood as the estimation of developers in SourceForge, identified<sup>7</sup> as Croatian. The 104 projects in SourceForge that are identified as Croatian-driven have a majority (i.e. more than 50%) of Croatian developers in their teams. Table 5.31 lists these numbers, including SourceForge for completeness.

Forge	Registered Users	Registered Projects
Linux.hr	ND	ND
Sourceforge	1,286	104

Table 5.31: Registered users and projects in Croatian forges (data April-May 2007). SourceForge has been included for completeness (data June, 2006).

It is important to point out that not all registered users are active developers in the forges. Many of them could register and never join a development project, for instance. Projects, as well, may not make use of all development-related tools offered by the forges (and therefore, some kinds of data sources, may offer no data at all). This fact is clearly revealed in the summary table 5.32, where the number of SCM (analysed/found) repositories, committers, commits, mailing lists, software releases and size of the software is given.

Forge	SCM repos	Committers	Commits	MailingLists	Releases	SLOC
Linux.hr	7/11	6	758	7	9	12,445
SourceForge	12/104	14	2,750	ND	ND	ND

Table 5.32: Information sources that could be extracted from Croatian forges (April-May 2007).

Figure 5.25 shows the number of SCM repositories, mailing lists and software releases identified and analysed.

<sup>7</sup>Methodology report - Chapter Methodology - Section Global forge's analysis

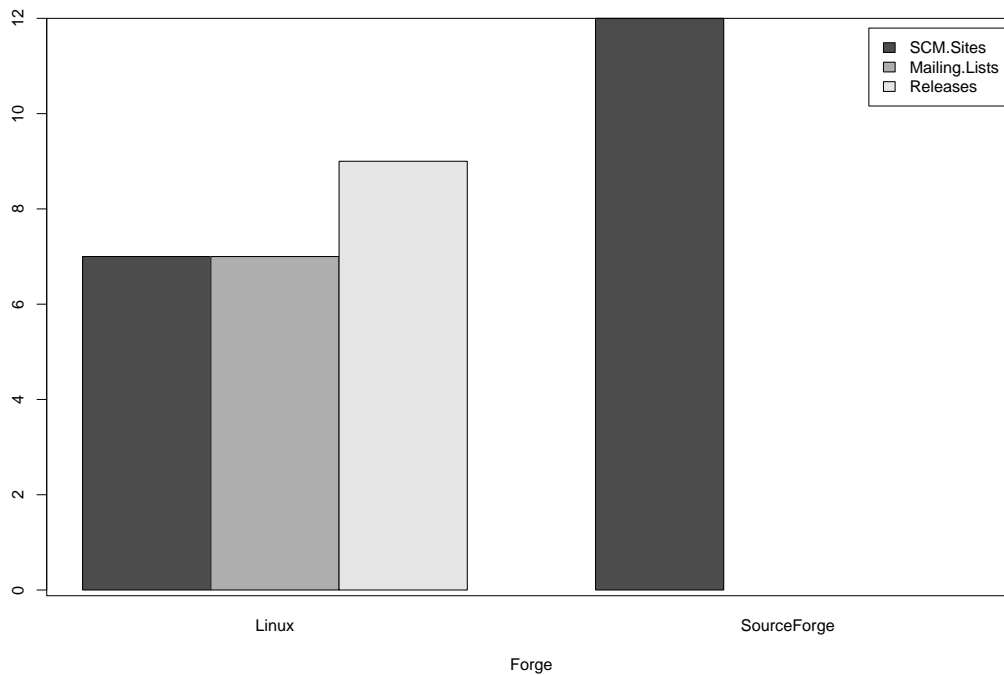


Figure 5.25: SCM repositories, mailing lists and software releases found in forges

## 5.7.2 Programming languages

The programming languages used for writing the code found in forges vary from forge to forge and from project to project. Table 5.33 and figure 5.26 contain information about the percentage of use of programming languages in *Linux.hr*.

Programming language	N. of detected lines	Percentage
C	10,980	88.2282
Perl	972	7.8103
SH	493	3.9614

Table 5.33: General Language results

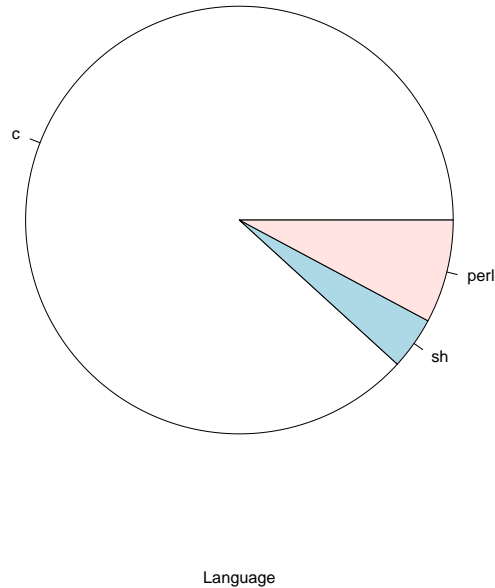


Figure 5.26: Programming languages in *Linx.hr*

### 5.7.3 Authorship

The authorsip data is based on the analysis of copyright notices in the source code. This analysis is performed on the release distributions found in the forges, using some heuristics for grouping authors (or, to be more precise, copyright holders) in some large categories. In this case, there is only one category. Just 6 different individual authors have been found. This data is consistent with the committers data, also 6 different people. Table 5.34 shows these results.

Type of author	Detected number
Individual Author	6

Table 5.34: Type of authorship in Croatian forges

### 5.7.4 Other data

In addition to analysing forges, the study tried to find some other sources of information that could help in the identification of community-related parameters, and could also be a source for further analysis. As a result, the data shown in table 5.35 was collected.

Region	Communities	Developers	Lugs	Projects
Croatia	2	23	3	3

Table 5.35: Other data collected

In this table *Communities* means group of users interested in libre software, *LUGs* are Oficial Linux User Groups (with a physical address) and *Platforms* are web sites which provide any kind of support to the libre software world, such as forges.

## 5.8 Bulgaria

The *Openfmi* forge, maintained by Sofia University, was the only local forge identified in Bulgaria during this study. This forge has almost 1000 registered users and 183 projects.

Besides *Openfmi*, this study has also analysed SourceForge, one of the most important worldwide development community, identifying more than 3600 users and 408 projects as Bulgarians.

The mailing lists hosted by the *Openfmi* forge, are not the main way of communication for Bulgarian developers. Only 11 lists were found with 80 posters and a reduced number of messages (835). Most of these messages are automatically generated with information about the commits performed during the development. Only 3 lists (including the automatic one) have more than 100 messages.

A total of 90 projects (including data from SourceForge and *Openfmi*) use Source Code Management systems. SourceForge's total number of projects hosted is more than twice the number of projects hosted in *Openfmi*. Despite this fact, only 9% (40) of the projects in SourceForge use Source Code Management systems. While in *Openfmi* this raises up to 27% (50). This is influenced by the fact that the number of developers and commits in *Openfmi* is larger than in SourceForge. *Openfmi* has 68 developers and 50,000 commits in contrast with 46 developers of SourceForge with 17,000 commits.

### 5.8.1 Forges analysed

The Bulgarian FLOSSWorld partners have identified some communities, LUGs, forges and projects. Openfmi.net forge (as the only one forge detected) has been spidered in order to identify the projects it contains. Table 5.36 lists a relation of the number of projects and users registered in that forge. The world's most popular forge, SourceForge, has been added to the table as many Bulgarian developers and Bulgarian-driven projects have been found there. The number of registered users at SourceForge should be understood as the estimation of Bulgarian developers<sup>8</sup> in SourceForge. The 408 projects in SourceForge that are identified as Bulgarian-driven have a majority (i.e. more than 50%) of Bulgarian developers in their teams.

Forge	Registered Users	Projects
Openfmi	983	183
Sourceforge	3606	408

Table 5.36: Registered users and projects in Bulgaria forges (data 24th April 2007). SourceForge has been included for completeness (data June, 2006).

It is important to point out that not all registered users are active developers in the forges. Many of them could register and never join a development project, for instance. Projects, as well, may not make use of all development-related tools offered by the forges (and therefore, some kinds of data sources, may offer no data at all). This fact is clearly revealed in the summary table 5.37. This table ( 5.37) shows the number of SCM (CVS/SVN) repositories, committers, commits, mailing lists, software releases and size of the software.

Forge	SCM repos	Committers	Commits	MailingLists	Releases	SLOC
Openfmi	50	68	51,293	11	39	203,946
SourceForge	40	46	17,359	ND	ND	ND

Table 5.37: Information sources that could be extracted from Bulgarian forges (April-May 2007).

Figure 5.27 and figure 5.28 show number of detected committers and commits in Openfmi and SourceForge.

<sup>8</sup>Methodology report - Chapter Methodology - Section Global forge's analysis

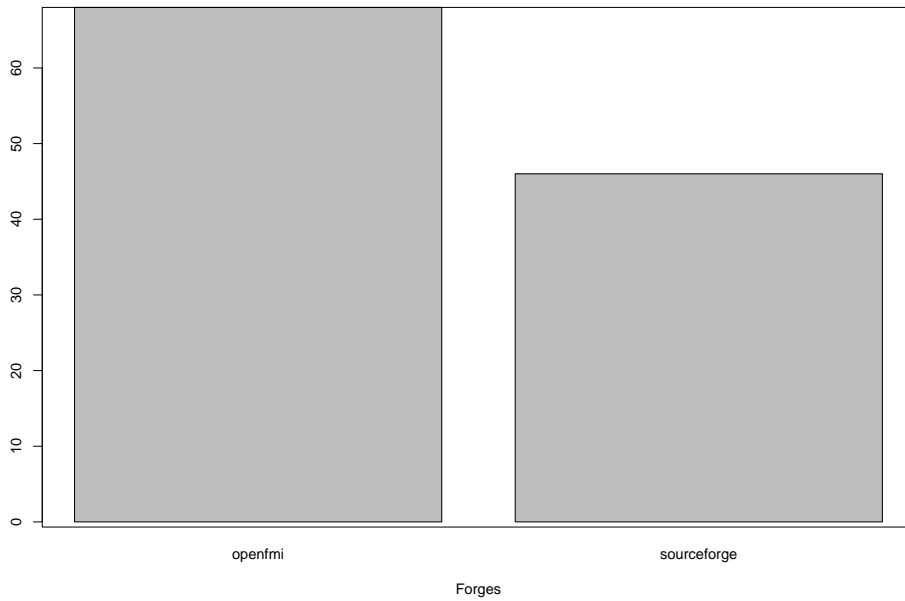


Figure 5.27: Committers per forge

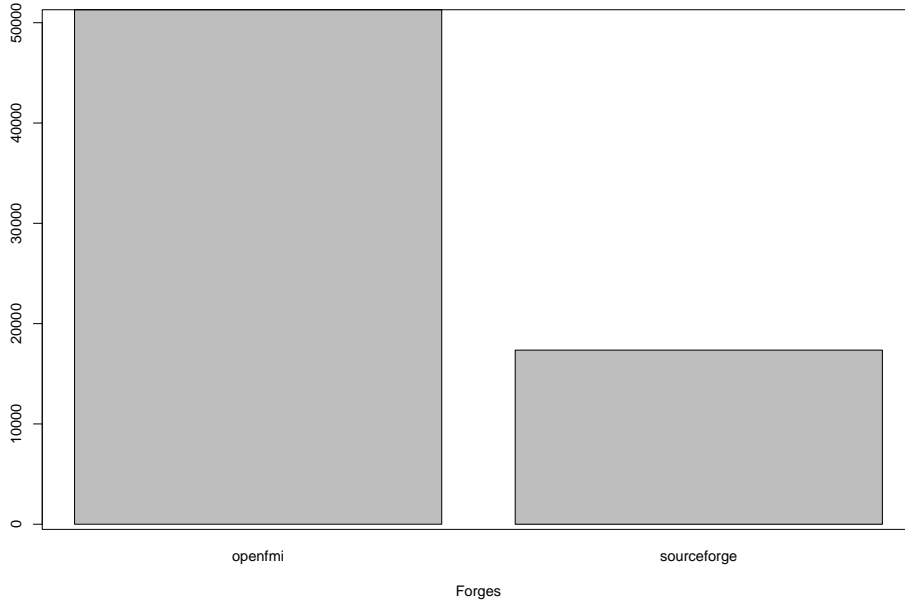


Figure 5.28: Commits per forge

## 5.8.2 Programming languages

The programming languages used for writing the code found in forges vary from forge to forge and from project to project. Table 5.38 and figure 5.29 contain information about the percentage of use of

programming languages in *Openfmi*.

Programming language	N. of detected lines	Percentage
c++	77574	38.0365
php	46610	22.8540
ansic	31065	15.2320
sh	29865	10.0913
java	7487	3.6711
perl	7342	3.6000
jsp	2950	1.4464
pascal	770	0.3775
python	245	10.0913
asm	35	0.1201
ruby	3	0.0014
Total	203946	100.0

Table 5.38: Programming languages used in Openfmi forge

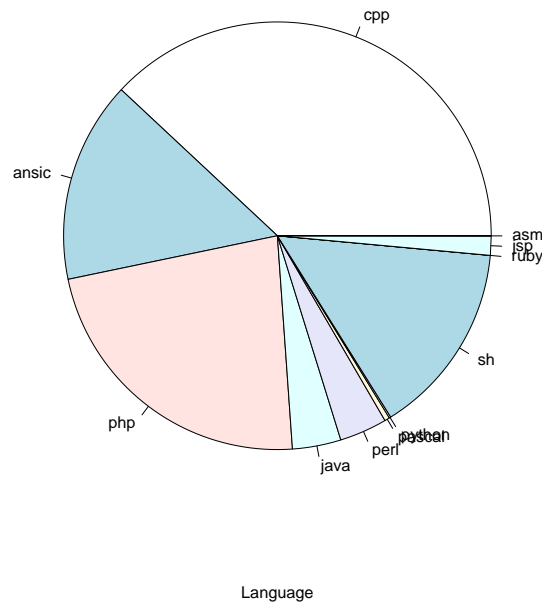


Figure 5.29: Programming languages used in Openfmi forge

### 5.8.3 Other data

In addition to analysing forges, the study tried to find some other sources of information that could help in the identification of community-related parameters, and could also be a source for further analysis. As a result, the data shown in table 5.39 was collected.

<b>Region</b>	<b>Communities</b>	<b>Developers</b>	<b>Lugs</b>	<b>Projects</b>	<b>Platforms</b>	<b>Forges</b>
Bulgaria	6	15	1	31	1	1

Table 5.39: Other data collected

In this table *Communities* means group of users interested in libre software, *LUGs* are Oficial Linux User Groups (with a physical address), and *Platforms* are web sites which provide any kind of support to the libre software world.