SIXTH FRAMEWORK PROGRAMME

PRIORITY 2.2.5

SOFTWARE AND SERVICES

Contract for:

SPECIFIC SUPPORT ACTION

Annex I - "Description of work" (public abridged version)

Project acronym: FLOSSMETRICS
Project full title:
Free/Libre/Open Source Software – Metrics and Benchmarking Study
Proposal/Contract no.: 033982
Related to other Contract no.:

Date of preparation of Annex I: 12.04.2006

Start date of contract: 01.09.2006

This is the public version of the Description of Work of the FLOSSMETRICS project. It has been stripped from some confidential information, but otherwise, it is almost a copy of the original document, version 2.0.
The version identifier of this document (2.0-public) reflects this origin.
More information about the FLOSSMETRICS project can be found at http://flossmetrics.org

Table of contents

1. Project Summary....4
2. Project objectives and state of the art....5
Current state of the art....5
Enhancing the available information....6
Main objectives of the project....7
Tasks to be performed....7
Measurable indicators....8
3. Participants list....9
4. Relevance to the objectives of the IST Priority....10
5. Potential impact....13
5.1 Contributions to standards....16
5.2 Contribution to policy developments....17
5.3 Risk assessment and related communication strategy....17
6. Project Management and exploitation/dissemination plan....18
6.1. Project management....18
6.2 Plan for using and disseminating knowledge....21
6.3 Raising public participation and awareness....22
7. Workplan....23
7.1 Introduction....23
7.2. Work planning and timetable....30
7.3. Graphical presentation of work packages....31
7.4. Workpackage List....33
7.5. Deliverables List....34
7.6. Work package Descriptions....36
8. Project resources and budget overview....48
9. Other issues....49
9.1 Ethical Issues....49
9.2 Gender issues....49
9.3 Concertation activities....50
9.4 Roadmap....51

1. Project Summary

Industry, SMEs, public administrations and individuals are increasingly relying on libre (free, open source) software as a competitive advantage in the globalising, service-oriented software economy. But they need detailed, reliable and complete information about libre software, specifically about its development process, its productivity and the quality of its results. They need to know how to benchmark individual projects against the general level. And they need to know how to learn from, and adapt, the methods of collaborative, distributed, agile development found in libre software to their own development processes, especially within industry.

FLOSSMETRICS addresses those needs by analysing a large quantity (thousands) of libre software projects, using already proven techniques and tools. This analysis will provide detailed quantitative data about the development process, development actors, and developed artifacts of those projects, their evolution over time, and benchmarking parameters to compare projects. Several aspects of libre software development (software evolution, human resources coordination, effort estimation, productivity, quality, etc.) will be studied in detail.

The main results of FLOSSMETRICS will be: a huge database with factual details about all the studied projects; some higher level analysis and studies which will help to understand how libre software is actually developed; and a sustainable platform for continued, publicly available benchmarking and analysis beyond the lifetime of this project. With these results, European industry, SMEs, as well as public administrations and individuals will be able to take informed decisions about how to benefit from the competitive advantage of libre software, either as a development process or in the evaluation and choosing of individual software applications. The project methodologies and findings go well beyond libre software with implications for evolution, productivity and development processes in software and services in general.

2. Project objectives and state of the art

The main objective of FLOSSMETRICS is to construct, publish and analyse a large scale database with information and metrics about libre software development coming from several thousands of software projects, using existing methodologies, and tools already developed. The project will also provide a public platform for validation and industrial exploitation of results.

In recent years, libre (free, open source)[1] software has developed as a novel form of collaborative production. Since its origin as a collaboration between individual volunteers, it has seen tremendous success, both in terms of the commercial and technical strengths of the produced software itself, but also as a model of organisation and development: open source software is arguably one of the best examples of open, distributed models for production and development that exists today. What is more important, from the point of view of the classical approaches to development methodologies by groups of professionals (and specifically from the point of view of the classical concepts of software engineering), the models used in libre software development are innovative in several ways, to the point that they are only recognized as valid models at all since they have actually produced mature and stable software: any previous "theoretical" analysis would have probably concluded that libre software development was not capable of producing any sustained, useful output.

In this context, FLOSSMETRICS will analyze in depth, from a quantitative point of view, a large quantity of projects, using mainly publicly available data sources. This analysis will help to better understand the landscape of libre software development, and to obtain factual data about it which can be used to improve libre software development itself (be it done in volunteer or corporate contexts), and to identify interesting practices that could be used in other contexts, but also to obtain indicators and data useful for companies willing to use libre software, or for public administrations intertested in its promotion or adoption. In addition, a huge database with quantitative data about thousands of libre software projects will be made available for the use of other research groups, what hopefully will act as a motivator to increase the empirical research on libre software development (and on software development in general).

Current state of the art

Despite the recent advances on the study of libre software development models, there is no such thing as a manual of libre software development (a document which would compile good practices, identify common known problems needing to be addressed, success strategies, etc.), and we are still very far away from that point. We only have detailed systematic descriptions of just a few of the libre software projects which have produced useful software packages, and a lot of gaps continue to exist in our knowledge of how the role of developers, external contributors, their structure and organisation, etc, affect the output of a libre software project, its productivity and growth, and on the development model used by it.

[1]In this description of work we will use the terms "libre software" to refer both to "free software" and "open source software", according to the corresponding definitions by the Free Software Foundation and Open Source Initiative, thus avoiding the connotations of both terms, which are rejected each by many actors in the community. "Libre" is well understood by most English speaking people, doesn't have the meaning "gratis", and is quite natural for romance languages speakers. Because of these reasons, it has been used in the past, specially in European contexts and more widely in the acronym FLOSS.

The FLOSS project (funded by IST/FP5) resulted in the single largest knowledge base on open source usage and development worldwide, and filled some of these gaps, at least in our understanding of the economic and software development models behind open source. Of particular relevance, the FLOSS project included a quantitative authorship analysis of 25,000 open source projects (the largest database of authorship information in libre software projects to date). In the context of the CALIBRE coordinated action (funded by IST/FP6) some actions are being taken towards getting quantitative data from a large collection of libre software projects[2] (forming probably the most comprehensive database to date with some development-related information about tens of projects, and raw information from CVS repositories from thousands of projects). These studies databases, and the studies performed on them, have helped to better understand libre software development.

However, those results can be considered as starting points in the field quantitative data collections about libre software development: a wide field is still uncharted in relationship with the massive retrieval of data from public development repositories, and there exists still little knowledge in the form of empirical, proven facts regarding open source projects – what forms of organisation and development exist, what is the relationship between organisation quality and sustainability, what forms of development are the most reliable. Indeed, even simpler questions – how do developers interact, how dependent are projects on one another, how responsive are they to bug fixing and other user requirements, how productive are developers, really – have no objective answers.

Other studies do exist on libre software development, but they are usually devoted to only a handful of projects. Of these, the studies on the Linux kernel, GNOME, Apache and Mozilla are the more detailed and known. Although they miss the larger landscape of the whole world of libre software development, they do show interesting details of the development model of the specific projects studied, and of the resulting software products. However, only with reliable data about a large quantity of projects would it be possible to start venturing sound theories about libre software development models, their advantages and problems, the identification of best practices and success strategies. In other words, until we have such data, with enough detail, about the historic evolution of a large number of projects, we can expect little advance on our understanding of this new way of developing software.

Enhancing the available information

FLOSSMETRICS aims to develop a base of such data, and build upon it by applying analytical models and techniques. Fortunately enough, libre software projects are known for their open development processes, during which huge quantities of information about the project are made available in the Internet, in many cases in data formats simple to retrieve and parse by using automated tools. For instance, version control systems (extensively used in libre software projects since several years ago) provide very detailed information about who was doing what and when, and about the historic evolution of the source code itself. Public mailing lists and forums provide a lot of information about the communication channels used in the system, and about the decision making process. Bug tracking systems provide details about the problems found with the software, and the way they are solved. All these (and more) data can be retrieved, stored and later analysed in an automated way, with frequent and continuous (also automated) monitoring and updates. Once the data for a large quantity of projects is available in comparable conditions, comparative and statistical analysis can be

[2]Some preliminar results, provided by URJC, one of the partners in this projects, are already available at http://libresoft.urjc.es

performed in a scale completely without precedent in the history of software engineering. The available data will also be used to calibrate simulation models of libre software development, which will be a valuable tool in order to predict the future of the development or to perform various "what if" scenario analysis testing different libre software management tactics.

Furthermore, in terms of productivity studies and cost/effort estimation models for software development, industry-popular models such as the COCOMO and COCOMO II series[3] and other cost/effort estimation techniques are hampered by their dependence on data under non-disclosure agreements. The information obtained by the FLOSSMETRICS project would significantly enhance the available information, based on empirical data far greater than the empirical basis for any (proprietary) software similar study that exists today.

Main objectives of the project

The main aim of FLOSSMETRICS is to construct, publish and analyse a large scale database with information and metrics about libre software development. Using existing methodologies and tools developed by members of the consortium, we will perform quantitative analysis of previously performed only on several dozen projects on several thousand software projects, for the first time allowing analysis and benchmarking based on robust large scale evidence. FLOSSMETRICS will provide a public platform for validation and industrial exploitation of results.

The main focus of this project is on the development activities in libre software projects, including the actors (developers), the artifacts (including the produced code) and the processes. In the short term, a complete, up-to-date view of the situation of thousands of libre software projects, and some of their interrelations will be provided. In the long term, the studies made possible by the data obtained will allow the identification of techniques and procedures for estimating the future of a project with a certain probability, when its past is compared with similar projects (which would be of great importance for a libre software project leader or core developing team). An additional goal is to get enough data to measure the productivity and development rate, with the final aim of getting accurate estimators for libre software projects.

Tasks to be performed

The project will perform the following tasks (details are in the work package descriptions):

- Identify and evaluate sources of data and develop a comprehensive database structure, built upon the results of CALIBRE (WP1, WP2)
- Integrate already available tools to extract and process such data into a complete platform (WP2)
- Build and maintain an updated empirical database applying extraction tools to thousands of open source projects (WP3)
- Develop visualisation methods and analytical studies, especially relating to benchmarking, identification of best practices, measuring and predicting success and failure of projects, productivity measurement, simulation and cost/effort estimation (WP4, WP5, WP6, WP11)

[3] http://sunset.usc.edu/research/COCOMOII/ resp. Boehm, B.W. (1981). *Software Engineering Economics*, Englewood Cliffs, N.J.: Prentice-Hall; and Boehm, B.W., Abts, C., Brown, A.W., Chulani, S., Clark, B.K., Horowitz, E., Madachy, R., Reifer, D.J. & Steece, B. (2000). *Software Cost Estimation with COCOMO II*, Upper Saddle River, N.J.: Prentice Hall.

- Disseminate the results, including data, methods and software (WP7)
- Provide for exploitation of the results by producing an exploitation plan, validated with the project participants from industry especially from an SME perspective (WP8, WP9, WP10)

Measurable indicators

The FLOSSMETRICS project workplan is designed to ensure both the fulfilment of the goals of the project and the monitoring of its success. In addition to the list of deliverables that allow for a detailed monitoring and evaluation of the output, some additional indicators are defined below:

Measurable indicators of success (over FLOSSMETRICS project duration)	**Count (minimum)**
Total number of libre software projects studied	5,000
Total number of workshops established by participants	7
Total number of academic publications and presentations related to the project results	15
Total number of citations in the trade press or media	20
Number of visits (lasting less than a week) to FLOSSMETRICS dissemination website (*)	100,000
Number of SMEs and industry representatives contributing to exploitation plan	15

(*) Number of visits to the website will be evaluated as the number of distinct (for a period of a week) visitors (estimated through the use of cookies, unique IP addresses, or similar techniques). That is, if the same visitor visits the site more than once during the same week, that would be counted as just one visit. But if the same visitor visits the site twice, in different weeks, that would be counted as two visits.

3. Participants list

Partic. Role	*Partic. No.*	*Participant name*	*Participant short name*	*Country*	*Date enter project*	*Date exit project*
CO	1	Universidad Rey Juan Carlos	URJC	ES	Month 1	Month 30
CR	2	University of Maastrich	UM	NL	Month 1	Month 30
CR	3	Wirtschaftsuniversitaet Wien	WUW	AT	Month 1	Month 30
CR	4	Aristotle University of Thessaloniki	AUTH	GR	Month 1	Month 30
CR	5	Conecta s.r.l.	Conecta	IT	Month 1	Month 30
CR	6	Zea Partners	ZEA	BE	Month 1	Month 30
CR	7	Philips Medical Systems Nederland B.V.	PMS	NT	Month 1	Month 30

4. Relevance to the objectives of the IST Priority

This project directly supports the goals of IST Priority Objective 2.5.5, "*Software and services*". This objective aims to "*support the competitive position of European software industry (notably SMEs) in more globalised and service-oriented markets*", by "*creating and extending open and interoperable platforms, methodologies, middleware, standards and tools*" and "*the simple and lowcost creation of new types of services and applications*". Europe leads the world in the development of open source software[4], and a better exploitation of this leadership by European industry is key to Europe's position in a globalised market. Open source software is one of the main driving forces behind the shift from product to service-orientation in the software industry today, and exploitation of this by European industry and SMEs will be key to their competitiveness. The low-cost rapid-prototyping made possible by some open source technologies such as Plone, Zope, and scripting languages like PHP and Perl are one reason new services and applications are often open source based. Indeed, most leading service-oriented commercial sites worldwide (Google and Amazon being the best known) are strongly based on libre software technologies, due to some of their intrinsic properties, such as reliability, adaptability, efficiency, etc.

This project supports the specific focus areas 3, 4 and 5 of the Priority Objective 2.5.5. It is crucial for "*Research into technologies specifically supporting the development, deployment, evolution and benchmarking of open source software*" through the extensive benchmarking and long-term evolution tracking (over 10 years) of open source software. It will provide data and analysis to support Focus area 3, "*Investigation into the use of open source models for improving software engineering*". The proposed action develops, implements and collects extensive measurements of "*agreed indicators of productivity and quality*" and through its study of productivity – substitution cost as well as real cost of open source codebases – will "*result in a measurement of the economic impact of OSS*".

The data and analysis will support Focus area 4, "*Foundational and applied research to enable the creation of software systems with properties such as self-adaptability, flexibility, robustness, dependability and evolvability*", all of which are claimed to be among the advantages of open source software development. Through the automated analysis of source code and other artefacts of FLOSS development, (WP1-4) as well as through the analytical studies (WP5, 11), FLOSSMETRICS will address the focus on "*high level methods and concepts ... for system design*". In particular, through the analysis of dependencies between projects at the code, module and developer levels and the user feedback mechanism, the project addresses in detail the topics of "*collaborative and distributed development; end-user development*".

And as a Specific Support Action, this project will of course support Focus area 5, "*Support actions contributing to the achievement of this strategic objective*". As all the results will be publicly available, including the databases and knowledge sets made available in interoperable open standard formats, it will allow synergies with other research including RTD projects and "*should help reduce fragmentation of research effort and build a critical mass of support for consensual action and agenda-setting*".

FLOSSMETRICS will provide for the "*widest possible use of results may be promoted through the use, extension ... of open source software*", and the "*use, extension and creation of open standards*" specifically for the interoperable exchange of metadata on software quality

[4]In terms of share of all developers, according to the EU (FP5) FLOSS survey as well as the Stanford FLOSS-US survey

and benchmarking. The consortium involves "*strong industrial users* [PMS] *join forces with software and service suppliers* [ZEA and Conecta, both SMEs] *in building common platforms and applications* [benchmarking, quality and productivity metrics] *with support of academic research partners"*. It supports indirectly the goal of "*International cooperation, notably with China or India, especially in the field of free and open source software"* as the lead partners UM/MERIT and URJC are both leading the on-going FP6 FLOSSWorld project which includes studies of free/open source software development techniques and differences among consortium member countries: the consortium includes the top software research and industry organisations in India, China, Brazil, Malaysia, South Africa, along with external support from Japan and USA (Stanford). The infrastructure and benchmarking facilities provided by FLOSSMETRICS will be easily implemented in these countries through the partnerships developed in FLOSSWorld.

An improvement in the development process of projects is expected to arise from various enabling factors resulting from the research carried out in the FLOSSMETRICS: the ability to measure development activity in detail and monitor it in comparison to past performance and to other projects; the ability to measure, categorise and benchmark projects based on the (socio-economic) organisational structure of their contributing developers as well as the (software engineering) organisational structure in terms of dependencies between software components. The FLOSSMETRICS predictive models, identification of best practices, and cost/effort estimation model would also clearly provide a platform for improvement of open source development.

Although not a development environment in itself, the tools integrated and knowledge base developed by FLOSSMETRICS, together with the interface and visualisation systems (WP6) will form an integral input to any future open source project's "*Open and modular development environment"* by providing a continually updated reference (WP3,4,5) against which a project's technical, organisational and structural progress can be measured and benchmarked. This will also provide support to Focus area 2 of Objective 2.5.5, allowing the evidence-based development of "*Principles, methodologies and tools for design, management and simulation of complex software systems"*.

As an SSA, FLOSSMETRICS also adds special value for research from an industrial perspective. Although the main parts of the research are being conducted by University members of the consortium, the strong interest and involvement from the industrial/SME partners ZEA and Conecta is clear. Furthermore, the interest expressed in the FLOSSMETRICS project by industry, specifically by PMS as a partner, proves the value of the research in this project to industry. As can be seen by the roles envisaged for collaboration with industry (see section 6.2 and workpackage WP8-10) the results and ongoing research and technology development of the FLOSSMETRICS project is expected to lead to significant industrial and SME exploitation of the results, well beyond the participants in this project.

Currently, the EU already leads in the usage and development of open source software, and after the FLOSS and other projects it is also a leader in research in this new field. The FLOSSMETRICS project aims to maintain this lead in a competitive global environment, especially in the domain of next generation methodologies for software engineering and development, and the growth of a globalised and service-oriented industry in both primary and secondary software sectors.

It is also important to notice that this project is conceived as a seed for further analysis, and for the set up of a permanent infrastructure of libre software projects analysis (something along the lines of an automated libre software observatory). In the best libre software tradition, all integrated tools and results will be released as libre (free, open source) software, and all the data will be made available in an easy to search and retrieve way. It is expected that these actions, alongside the more classical dissemination activities will help to make the results of the project a de-facto standard on the libre software engineering world. The interest of PMS in particular, which plans to use the tools and benchmarking system also to study the efficiency of its internal (proprietary) software development shows that the results are certainly not limited in relevance to open source software, but are highly relevant to the classical, proprietary software engineering world and the full spectrum of the software industry.

FLOSSMETRICS will also provide support to IST Priority Objective 2.5.8, "*ICT for Networked Businesses"* in particular for Focus area 1, "*providing an open-source environment and suitable operative models enabling small- and medium-sized organisations to cooperate*". For successful operation in an open-source environment it is crucial for SMEs to know what makes open source-based cooperation - "*dynamic virtual organisations"* - function effectively. SMEs also need to know whether such cooperation is viable. FLOSSMETRICS's benchmarking and productivity data provides insights into the economic sustainability of such efforts, and WP8 (led by Conecta, an SME that works in an open source environment) will examine the implications for SMEs in particular.

Finally, while not in this call, FLOSSMETRICS also supports the goals of IST Priority Objective 2.3.2.3 from the IST Workprogramme, "*Open development Platforms for software and services*". This objective aims to "*build open development ... environments for software and services providing the next generation of methodologies, ... and tools to support developers ... in the production of networked and distributed software systems*." Open source software is the most widely used production system for networked and distributed software development. FLOSSMETRICS aims to build tools and apply them in a publicly accessible and widely disseminated (WP7) open environment to facilitate the improvement of software development in open source projects, thus clearly meeting this objective of the IST Priority.

5. Potential impact

Proven previous impact

The IST/FP5 FLOSS project had tremendous impact. Since the release of the final report in June 2002, there have been over 15,000 visits a month to the FLOSS web site and report. "FLOSS", created as a project acronym, has now become a widely accepted generic term bridging the confusing gap between the terms Free (or Libre) Software and Open Source, and has even been recommended as a generic term by people like Richard Stallman, founder of the Free Software Foundation. A search on Google shows over 75,000 websites referring to the FLOSS project[5] and several mirror sites duplicating the final report in its entirety. The project findings were covered by most major industry publications worldwide and press reports in several languages. The FLOSS team has proved successful at following up on the experience of the FLOSS project, with the acceptance of the successor FLOSSPOLS project as among the first contracts to be signed under the IST FP6. Developers were the most interested in the FLOSS results – the developer survey and quantitative authorship analysis of over 25,000 open source projects were the most downloaded parts of the FLOSS report. Similarly, the Orbiten Survey of source code authorship of 5,000 open source projects – the first of its kind when released in 2000 – had 100,000 downloads within a few days. The FLOSSWorld press release led to 50 000 links on Google (and over 100 press articles in over 30 languages) within 3 weeks of its publication, in the second month of the project.

FLOSSMETRICS is expected to have no less an impact within the community of developers, users, industry and those who study the libre software phenomenon, since the issues addressed – especially in the depth and scope proposed – have rarely been examined before, and certainly not in terms of the proposed comprehensive extraction and analysis of empirical data. The FLOSSMETRICS project will have a broad impact on the academic research and policy communities through the sustainable, public and open design of the planned empirical database, providing a continually replenished source of further research well beyond the end of the project. Similarly, this sustainability will provide a continuing support environment for open source developers.

Impact in the software domain

The impact of the project is expected to be large in the libre software development realm (and probably in the whole software development landscape). First of all, FLOSSMETRICS is expected to produce the most complete and detailed view of the current landscape of libre software development currently available, providing not only a static snapshot of how libre software projects are performing now, but also (in some aspects) providing historical information about the last ten years of libre software development. In addition, since all of this will be in the form of quantitative data made available to other researchers, it is expected that the project, and the methodologies and data collections produced by it, become a mandatory reference for further studies in the field.

From the software development point of view, the quantity of data available about how libre software is developed will also be of great value, both as a contrast for studies in the proprietary software world, and as data about software on which to test and check any generic model about software development. The dataset produced by the project will be not only the largest in the context of libre software development, but for sure one of the largest and more

[5] Search term: +FLOSS software OR open OR source OR libre OR free -dental -site:www.infonomics.nl -site:floss1.infonomics.nl

comprehensive in the whole field of software development. This will allow researchers especially from software engineering to test several assumptions and models currently debated, including new software development paradigms like aspect-oriented development.

Exploitation strategy

The results of the FLOSSMETRICS project are expected to be very useful to all members of the consortium. For the University members (consortium partners 1 to 4) those results clearly form the basis for further research, and also a basis for further research by other members of the academic community. In particular, the database and studies developed within the project will provide an excellent basis for further research and policy studies – the main form of non-commercial research exploitation for any public university. Although all research results and databases will be public and tools used are publicly available libre software programs already developed (mostly by members of the consortium), the FLOSSMETRICS consortium partners involved in their creation clearly have an advantage in terms of first and most extensive experience with the database, research results and usage of tools in this large scale. Furthermore, there is a clear strategic potential for exploitation by the libre software community and libre software industry in general. This potential will be identified and detailed in the exploitation report in WP10, based on the validation and feedback from industry.

For industrial players, FLOSSMETRICS is expected to provide significant insights into the function and organisation of libre software projects. The system which will be set up for automating the data retrieval and low-level analysis will be useful for the continuous monitoring of the organisational structure and productivity of the libre software projects in which industrial developers are involved, as well as the study of the code structure itself - especially the aspects of code reuse and inter-dependency among modules. Specifically, PMS, being a partner of the FLOSSMETRICS consortium, has expressed a strong interest in the research results of FLOSSMETRICS, and is interested in collaboration on validation and exploitation (WP9).

The database of metrics on open source projects, together with the methodologies of the studies produced in FLOSSMETRICS will be useful for the industrial partners to compare the productivity and efficiency of its open source developer teams with past performance, with each other, and with libre software projects developed outside, in other companies or in the open source community in general. PMS and ZEA are particularly interested in collaborating with FLOSSMETRICS with regards to the measurement of productivity and quality aspects of libre software development. Our analysis has previously identified over 1000 firms contributing source code to libre software projects (accounting for about 15% of all libre software code written). While firms are rarely able to value the extent of their contribution or monitor productivity, these are substantial and can be estimated, as shown in the table below (notably, the top contributor is an European SME). The diversity and extent of involvement in open source projects by industry – together with their current inability to precisely measure the value of their own output - demonstrates the impact of FLOSSMETRICS results on industry in general, well beyond the project participants.

Rank	*Name*	*Dev. Time (months)*	*Person-months*	*Cost (mil euro)*
1	Trolltech AS	88	11866	111
2	Sun Microsystems	73	7142	67
3	IBM Corp.	72	6997	66
4	Digital Equipment Corp.	68	5903	55
5	Silicon Graphics	67	5766	54

Copyright © 2005 URJC. Shows (rough) cost estimate for contributions to Debian, 2002.

The aspects related to code structure, modularity, code reuse and dependency will help industrial partners to compare the programming methods and inter-dependency of its open source developer teams, once again monitoring dynamic changes and in relation to other projects. The predictive and benchmarking studies will allow industrial partners to provide inputs to their developers in order to restructure teams, fill in gaps in organisational (developer) structure and modularity (code structure). This will improve efficiency and competitiveness by matching best practises in libre software development identified through metrics across several open source projects. The cost/effort estimation studies for libre software projects will help industrial partners better estimate costs for future libre software developments (by comparing with cases in the studies),and the methodolgies of such studies would help to better calculate gross and net productivity for libre software projects that industrial partners' developers may participate in. In addition, ZEA has expressed particular interested in monitoring the extent to which libre software projects reuse and link to code developed by their associates.

For consortium Partner 5, Conecta, an SME that derives much of its income from reports and detailed analysis of specific FLOSS software packages, there are specific additional benefits relevant to SMEs in general. For Conecta, as for most SMEs involved in commercially exploiting open source software, the main business model is related to so-called "integration and installation" services, and the main problem associated with such services is the identification of the libre software packages most appropriate for customers' needs based on a long term perspective. The benchmarking and analytical results from FLOSSMETRICS are expected to significantly help in this process. In particular, the analytical data that result from FLOSSMETRICS will help in describing the added dimension of the "activity level" of a project (see WP8), and so will be particularly useful in providing Conecta's customers with an indication of which open source application may be more appropriate in their medium-to-long term strategy. Thus, it will be possible to advise customers to choose appropriately between, for instance, a finished package that has, however, no more active development and a younger project with an active support community.

For libre software developers (in companies doing libre software development, or volunteers), public access to the integrated system and data generated by this project will provide a consistent framework for comparing and benchmarking their own projects. Using already existing tools, the system will allow for getting detailed reports about the state and history of a project (given their data repositories). Project leaders will be able to compare the results for their projects with other similar ones in the database system maintained by FLOSSMETRICS. In addition, the studies on estimation models and visualization will be of great help when planning for the future of their project, or when auditing its current state. All these services could even lead to a pay-per-use scheme at some future stage (specially for interested companies), which would produce benefits for both companies and developers.

The above benefits, especially benchmarking and cost estimation, may also be applicable to proprietary development. Since the tools used are libre software, industrial partners can use them internally – and in many cases they will have applications within proprietary software development. Of course those tools could be used without having notice of FLOSSMETRICS, but the visibility that the project will give to them, and the methodologies for how to use them in combination (provided as a part of the corresponding studies) will be a great motivator and facilitator. PMS has expressed a particular interest in exploring this application of the FLOSSMETRICS results, as they plan (under the name "Inner Source") to develop open source development methodologies for in-house software development. Finally, the exploitation and validation reports to which industrial partners are to provide inputs (in WP8, 9) will demonstrate the potential exploitation of FLOSSMETRICS results within the libre software community and other industrial users and developers of open source software.

Dissemination strategy

FLOSSMETRICS aims to provide a deeper understanding of specific issues related to the role of open source software in the Strategic Objectives, and to disseminate this understanding, through interaction with the target community, as widely as possible. All FLOSSMETRICS deliverables are intended for public release as soon as possible (while addressing privacy and practical considerations). As described in the dissemination work packages (WP7), a target community comprising three groups has been determined:

1. the libre software developer community itself, the primary target group for FLOSSMETRICS.
2. the business community with a special focus on SMEs and secondary sector companies with interest in libre software development
3. the academic, policy and policy-making communities

European added-value and relations to international research activities

The FLOSSMETRICS project supports Europe-wide goals, and clearly needs to be conducted at a European level. The studies described in this DoW are unprecedented in scope and depth worldwide, and is essentially transnational in nature. There is insufficient expertise to carry out such R&D in any single EU state. Libre software is directly relevant to the European Research Area (ERA) not only due to Europe's lead in libre software development and deployment, but also as the model for collaborative R&D best suited for ERA as it can be rapidly adapted and disseminated. FLOSSMETRICS studies on productivity and development methods strongly support the aim of eEurope2005 to ensure that Europe is the "most advanced knowledge based economy by 2010". The FLOSS workshop on "Advancing the Research Agenda on Free/Open Source Software" (October 2002) jointly supported by the European Commission and the US National Science Foundation clearly saw the imperative need for research in quantitative analysis, empirical data and productivity. The transnational FLOSSMETRICS team has already collaborated with Stanford University on a study of the Linux Kernel and received funding from the US National Science Foundation for further research; FLOSSMETRICS will continue such cooperation while maintaining Europe's lead.

5.1 Contributions to standards

Since many open standards (such as TCP/IP, HTML/HTTP, DNS/BIND, many XML-related standards, etc.) result from and are maintained by libre software developers, studying the

processes and organisation of such developers is useful. In particular, the studies to be performed will identify the degree of influence and concentration of contribution to reference implementations of standards by different players. Projects that involve widely used standards – de facto or de jure – will be specifically included among the open source projects that are selected for benchmarking and detailed study in WP 3-6. In addition, the formats and analysing methodologies which will be provided as outputs of this project will also be offered as standards to the software engineering research community. Since there are few previous efforts involving benchmarking software on a large scale with a willingness to exchange data with others, there are few standards in this area. So we expect to help to set up standards at least in the following areas:

- Data formats for exchanging information about libre software (and in general, software) projects
- Parameters for characterization of libre software projects
- Storage systems specialized in the storage, query and retrieval of detailed information about software development
- Architecture and interfaces for the integration of tools devoted to the retrieval and analysis of public data about software projects

5.2 Contribution to policy developments

There are some other EC policy related issues of relevance to the FLOSSMETRICS project, beyond those already set out in the IST Priority action line and already addressed in this description of work, which are described below.

Engaging with actors beyond the consortium

One of the goals of this proposed project is to interact and involve the wider community of developers and industrial and SME participants. WP7 defines this, as does to some extent WP8 and WP9. Clearly the FLOSSMETRICS project is relevant outside the research area and consortium partners, who have demonstrated within the dissemination and exploitation goals a willing readiness to "engage with actors beyond the research to help spread awareness and knowledge and to explore the wider societal implications of the proposed work". In particular, this description of work shows (section 6.2, "Plan for using and disseminating knowledge") the specific interest of several companies (including Conecta, ZEA and PMS, all of them members of the consortium), who have expressed interest in engaging with the FLOSSMETRICS project.

Relevance to education

The results are also relevant to education and especially higher education in software engineering, as they can further develop the field of open source software engineering which has the special feature that students can immediately have practical development experience, unlike with proprietary software. URJC in particular already has a related teaching programme, to which FLOSSMETRICS would contribute.

5.3 Risk assessment and related communication strategy

No specific risks for society or citizens have been identified by the consortium in relationship with this project.

6. Project Management and exploitation/dissemination plan

6.1. Project management

Management structure

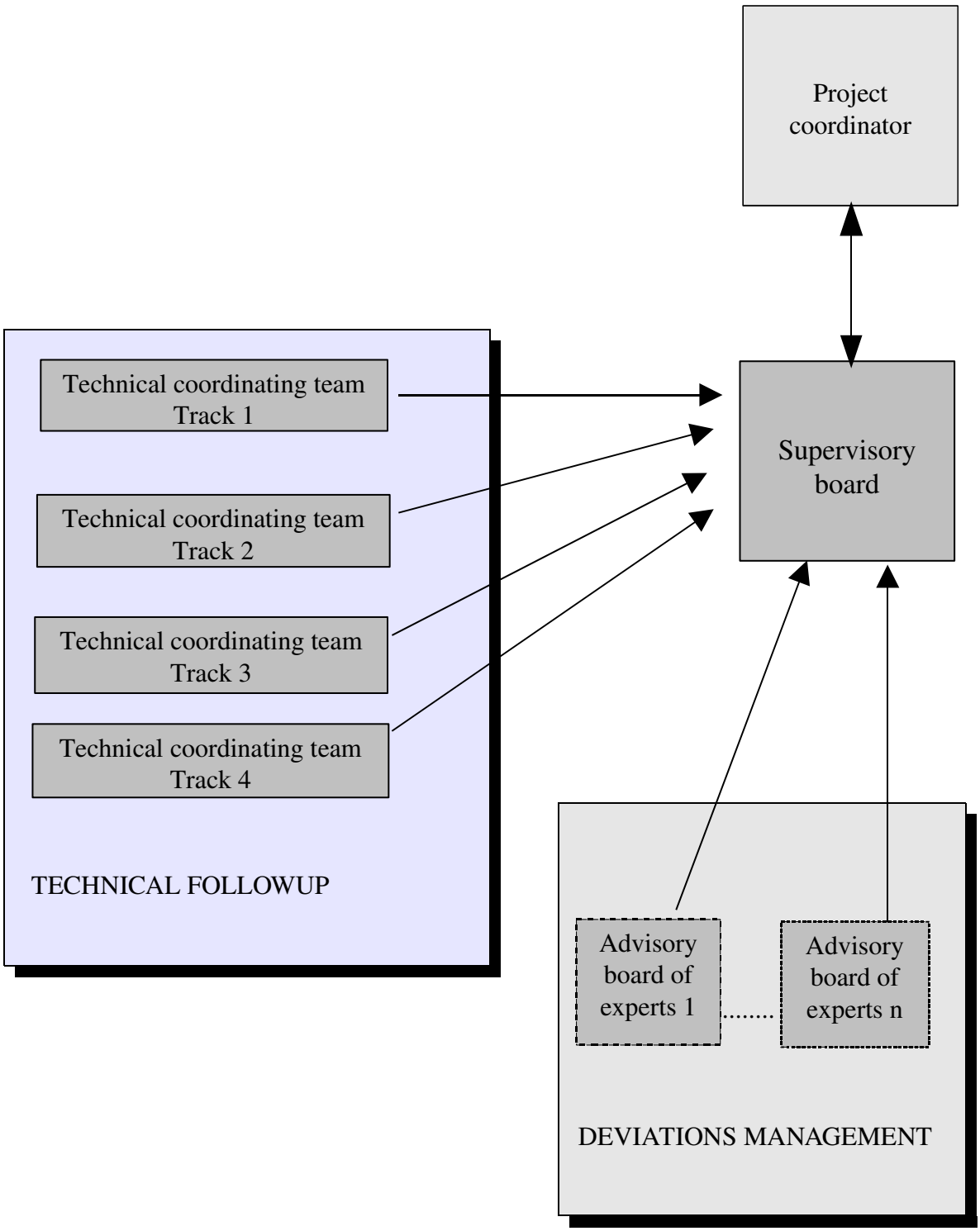
URJC will assume overall project coordination responsibilities. They will assume overall responsibility for monitoring the progress of the work packages, and ensuring that the objectives of the project are met; they will also be responsible for communications between the consortium and the Commission, the production of deliverables, the development of the exploitation plan (WP10). The project coordinator will be backed up by the general administrative structure of the URJC, which has experience in managing several large projects over many years, and one expert in managing projects, which will help in the micromanagement of the project, in preparing meetings of the consortium and in the relationships with the Commission.

Under this general management structure, individual experts participating in the consortium will assume responsibility for the day-to-day management of specific work packages. The organisation of the FLOSSMETRICS project is clearly split in four tracks, as described in the workplan (section 7). Each track will be coordinated by one of the partners (in bold, in the table below), with some other persons supporting it and leading work packages in that track. The coordinator and supporters (if any) for each track will form the technical coordinating team for that track, with responsibilities on the technical followup of the track, early detection of the deviations in that track with respect to the workplan, and the proposal (if needed) of corrective measures. The technical coordinating teams will interact on a regular basis through email (and if needed, through conference calls), and will meet (if needed) in coincidence with the meetings of the supervisory board (see below), although each team could organize their own meetings if they find them necessary. In particular, the team for Track 1, due to the complexity of the activities involved, will have at least two of those meetings or telematic conferences. In the case of Track 3 and Track 4, no coordinating teams are foreseen, and the partner coordinating those tracks will be the main responsible for the coordination.

Track (WPs)	Description	Coordinator (bold) and supporters
Track 1: 1,2,3,4,5,6,11	Empirical foundation and focused studies	**URJC,** UM, WUW, AUTH
Track 2: 8,9,10	Exploitation/validation	**Conecta,** ZEA, PMS, UM
Track 3: 7	Dissemination	**UM**
Track 4: 12	Management	**URJC**

A supervisory board will jointly monitor the progress of the project, based on the inputs from the technical coordinating teams for each track, and will ensure effective coordination between all the partners. This board will consist of representatives from all the partners, and will be chaired by the coordinator of the project, URJC. The board will maintain consultations on a regular basis and will meet up to three or four times during the project duration (about every 8-10 months), depending on the need felt at the time for closer coordination. In case of problems during the project implementation such as significant deviation from the workplan the board will consult to decide on alternative options and remedies, and if necessary the appointment of ad-hoc advisory boards of experts in case of unforeseen problems arising from research results or the methodology chosen. The members of such advisory boards will be chosen by consensus of the board, among the experts

proposed by the partners, based on their areas of expertise and the kind of problems, if and when this circumstances arise.



Summary graph of the management structure

Each work package will produce designated deliverables. At the end of the project, these deliverables, when in the form of papers, will be refined to produce a Publishable Final Summary Report, that will be written to be accessible to a broad audience. The integrated system put together during the project for the analysis of data (and based on already existing libre software tools) will be distributed through the project web site and elsewhere as libre software.

Management of knowledge and intellectual property

The knowledge and intellectual property resulting from the FLOSSMETRICS project will be in three forms:

- Data resulting from surveys and the application of data extraction tools. Data acquired will also be made publicly available while keeping in mind ethical concerns, specifically with regards to privacy.
- Papers and reports analysing these data, developing benchmarks or models, or reporting on the validation of the tools. Papers and reports are intended for widespread public dissemination, and will be licensed under Creative Commons Share-Alike licences when editorial policies of publishers allow for that.
- The system used to support the data retrieval and analysis. This system will be published under the GNU GPL license or another libre software license as appropriate.

No proprietary knowledge is expected to result directly from the FLOSSMETRICS project. In addition, some research on appropriate licenses for dissemination of the raw data will be done, with the aim of licensing those produced by the project in conditions similar to those offered by the GNU GPL or the Creative Commons Share-Alike family of licenses.

6.2 Plan for using and disseminating knowledge

The results of the FLOSSMETRICS project are expected to be very useful to all partners of the consortium. For the academic members, they clearly form the basis for further research, and also a basis for further research by third parties. The exploitation of the results from an industry, especially SME perspective, will be facilitated by the development of an exploitation plan in cooperation with the industrial partners (and specially Conecta and ZEA). This will allow for ongoing exploitation and validation of the results.

For Philips and other companies interested in exploring the novelties of libre software development (and specially those in the secondary sector), FLOSSMETRICS is expected to provide significant insights into the function and organisation of inner source projects. The retrieval and analysis system (which will be available for third parties, since it will be composed of libre software, including any glue-code developed, if needed, by the project) will be useful for the continuous analysis of the organisational structure and productivity of the inner source projects in which PMS' developers are involved, as well as the study of the code structure itself -especially the aspects of code reuse and inter-dependency among modules. PMS would be particularly interested in monitoring the extent to which inner source projects reuse and link to code developed within the company.

The database of metrics on inner source projects, together with the benchmarking and predictive models used in FLOSSMETRICS will be useful for companies to compare the productivity and efficiency of its inner source developer teams with past performance, with

each other, and with inner source projects developed by third parties. The aspects related to code structure, modularity, code reuse and dependency would help PMS and other companies to compare the programming methods and inter-dependency of its inner source developer teams, once again monitoring dynamic changes and in relation to other projects. The calibrated predictive and benchmarking models would allow PMS to provide inputs to its developers in order to restructure teams, fill in gaps in organisational (developer) structure and modularity (code structure). This would improve efficiency and competitiveness by matching best practises in inner source development identified through metrics across several inner source projects.

The cost estimation model replacing COCOMO for inner source projects would help companies to better estimate costs for inner source project development and calculate gross and net productivity for inner source projects that developers may participate in.

The above benefits, especially benchmarking and cost/effort estimation, may also be applicable to proprietary development. Since the tools are inner source, PMS and other companies can use them internally - many of the tools will have applications within proprietary software development.

Finally, the exploitation and validation reports will demonstrate the potential exploitation of FLOSSMETRICS results within the inner source community and other industrial users and developers of inner source software. The details of this exploitation and validation reports, and the details of the dissemination strategy, are described in subsection 7.1 (when describing the correspondent tracks) and in the description of the workpackages WP7-10 in subsection 7.6.

6.3 Raising public participation and awareness

In addition to the activities specifically targeted to the software industry (and the companies specifically interested in libre software), other dissemination activities will be performed. These activities will include communication to the media (including newspapers and magazines), and delivery of information in the website specifically targeted at raising public awareness. Some of these activities are scheduled in the context of concertation activities (see section 9.3).

7. Workplan

7.1 Introduction

FLOSSMETRICS aims at filling existing gaps in the knowledge about libre (free, open source) software projects, and at improving the understanding of the libre software development process so that its full power can be harnessed by industry. For this end, a large-scale empirical study will be undertaken, which will provide a base for facts-based analyses and studies. All of the results of this project (studies, analyses, including the methodologies used, and the data employed) will be made available publicly, in order to ensure the validity of the approach and to allow for continuous research, SME and industry feedback and exploitation sustainable well beyond the lifetime of the FLOSSMETRICS project, all of its results.

Some analytical studies using the resulting data will also be made to provide proven facts and results on the forms of organisation and development existing in libre projects, relationships between different variables like project size, number of participants, resulting quality and several others. In particular, the productivity of developers, and the efficiency of the different libre software development models will be analysed. This analysis will lead to benchmarking and identifying best practises, validation from the industry and SME perspective as well as from users and developers from the open source community, and result in an exploitation strategy for the results of this research.

To reach these results, FLOSSMETRICS can be broadly divided into four different tracks, each composed of one or more workpackages: empirical foundation and focused studies; dissemination; exploitation and validation; and management. The following tables summarize the workpackages (below) and tracks (after it, including the workpackages in each of them). After them, the tracks will be described in detail. For a graphical presentation of the tracks and respective workpackages, see section 7.3 (Graphical presentation of work packages).

Workpackage	Workpackage name	Track 1: (Empirical foundation & focused studies)	Track 2: (Dissemination)	Track 3: (Exploitation & validation)	Track 4: (Management)
WP1	Data Sources	X			
WP2	Retrieval syst. integration	X			
WP3	Database	X			
WP4	Analyses	X			
WP5	High level studies	X			
WP6	Visualisation	X			
WP7	Dissemination		X		
WP8	SME Exploit./Valid.			X	
WP9	Ind. Exploit./Valid.			X	
WP10	Exploitation plan			X	
WP11	Productivity	X			
WP12	Management				X

Table: Workpackages and tracks

The remainder of this section is organized by tracks, providing an overall picture of all the workpackages, and their relationships. The activities performed in the context of all these

packages do not include the development of tools, since all those to be used are already available (except for maybe some glue-code for putting them to work together) neither the research on new models (proven approaches and models will be used when needed). There are also no demonstration activities in the FLOSSMETRICS project.

Track 1: Empirical Foundation and focused studies

The workpackages WP1, WP2, WP3 and WP6 aim at laying the empirical foundation for the work. The relevant data sources existing (repositories with publicly available information about libre software projects) will be identified, analysed, and the most suitable ones will be chosen in WP1. The necessary retrieval tools will be selected and integrated if needed (most of the tools have been developed by the partners in the project, and therefore integration problems are expected to be minimum) and executed in WP2. This will result in a database containing the relevant data forming the basis for the focussed studies (WP3). To facilitate both exploitation and explorative studies, a separate work package (WP6) is devoted to use existing graphical user interfaces and visualisation tools on these data.

Three different workpackages (WP4, WP5 and WP11) will be devoted to analysing the data provided by the empirical foundation. The first work package of these (WP4) develops a classification scheme for open source projects by uncovering significant relationships between several project variables, both directly retrieved and newly computed. This also leads to an update of the database to include these results. Based on this, WP5 digs deeper into software engineering issues and, for the first time, tries to look into the future by using and calibrating already available benchmarking, prediction and simulation models for dealing with project progress, staffing, quality and several other project features. Several techniques, including for instance social network analysis and archaeological techniques (using version control data to analyse the age of individual lines of code – i.e. for how long have they survived in the codebase; high survival rates may indicate good quality, or perhaps low rate of change, depending on what other indicators are identified). The last work package in this group (WP11) takes a more economic-oriented viewpoint in order to study the effects of this development model on productivity, and further use cost/effort estimation models to better understand this issue in libre software development.

It is important to stress that, since the result of packages WP1, WP2 and WP3 will be a set of automated tools to feed the database and analyse it, and since a huge quantity of projects are candidates for fetching their development data, the whole database is expected to include information about thousands (probably tens of thousands) of software projects, in the largest coordinated effort to date to accumulate information about software development. From the work plan viewpoint, this will imply that, although the first version of the database system will include information about a significant quantity of projects of all kinds, during all the duration of the project the database will be growing, updating its information about those projects, and including more and more on its tracking list.

The rest of the description of this track will provide more details about some of the mentioned workpackages.

To start with, the data produced by the FLOSSMETRICS project will originate from many sources, following the flow of data in the libre software development process. These are in the form of *primary sources* and *secondary sources,* identified in WP1:

Primary sources: The process of libre software development results in source code, a major primary source of data. Furthermore, the development process uses a number of development aids and tools, such as source code management systems (CVS, Subversion, etc.), which result in considerable amounts of *metadata* that is publicly available. Mailing lists associated with specific open source projects are yet another source of primary data and metadata. Metadata is extracted from primary sources of through the use of extraction/integration tools (many of which exist, such as CVSanalY which provides statistical information based on raw CVS and Subversion data; DrJones which does archaeological analysis on CVS data; GluTheos which analyses snapshots of code; MailingListsStats, which analyses mailing lists archives; etc.[6]. Metadata may also be *generated* from, e.g. source code itself – through the use of tools to count source lines of code (SLOC) or some complexity metrics (CCCC, cyclo, etc.), or tools such as CODD, which computes developer contribution based on names and signatures in source code, and builds a graph of dependencies between source code modules.

Secondary sources: include data sources not directly involved in the development process. For instance, web search engines can help provide an estimation of the projects usage, which is one measure of its success; or provide information on the links between different projects, or different project teams. Other examples can be GPG key servers, which can provide information on who knows whom, and on different identities of developers.

Metadata may also be generated from these sources. These data and metadata will be extracted and integrated through the use of those tools (in WP2) into the FLOSSMETRICS database (WP3). Processed data from the database is subject to further analysis through analytical tools (e.g. to identify changes over time or identify clusters of projects or developers), benchmarking and classification (WP4). In addition, WP6 uses existing tools to explore ways of visualizing the data and relationships, which is a major problem in itself given the vast amount of data available.

For classification (included in WP4), several indicators could be used, some of them being the community size around the project, the timing of messages posted to mailing lists, the statistical distribution of answers by message, the response time per message, the project age (and the age of its project components), the turnover within development teams, the time distribution in bug fixing, ratios such as Source Lines of Code (SLOC) or complexity / bug reports, SLOC or complexity / community size, SLOC or complexity / mailing list archive size, etc. All these indicators will be available from information in the database system, and tools will be provided to calculate them (and integrate the results into the database). Once projects are classified according through different criteria, each project population will be studied, to highlight the benefits of having such a large knowledge database about software projects, in order to answer questions such as: what organisational structures and code profiles are usual in successful libre software projects, (and how many ways can we define and measure "success" for libre software), how can development effort be estimated, which best practices are common in successful projects, and which simulation models predict better the evolution of libre software projects.

In order to achieve these, the statistical techniques that may be appropriate for estimation model building include ordinary least square analysis and ANOVA. Furthermore existing simulation models developed at consortium partner AUTH will be tailored to act as a "generic libre software simulators" in different aspects of libre software development. This models will utilise data from the project itself and output the population size of programmers

[6] A more complete list is provided at http://libresoft.urjc.es/index.php?menu=Tools

working per task (code writing, debugging, testing etc.) and per project module (e.g. system files, applications, drivers), the lines of code (LOC) written, the defect density and the structural quality of code for the whole or parts of the project as functions of time. It will be useful to simulate project bottlenecks and provide with qualitative ("semi-quantitative") results on libre software development dynamics under various "what if" scenarios.

Additionaly the efficiency and the applicability of several software cost estimation models will be explored and assessed. A comparative evaluation of several traditional cost estimation models such as analogy based algorithms and regression models will be performed with other state-of-the-art classification algorithms from the field of Machine Learning such as Classification and Regression Trees (CART), Association Rules and Bayesian Belief Networks. Comparing the above algorithms we will have the chance to evaluate software cost point estimate techniques (regression models, analogy based estimation) with software estimation techniques that produce intervals.

All the above techniques will be applied on historical libre software productivity data in order to calibrate cost estimation models to the individuality of libre software. Target of the software cost estimation models will be, taking into consideration of the libre software attributes, to provide a complete estimation framework involving the duration of the project, the amount of effort needed along with a justification of the particular estimate. The results of the models will be compared with software cost estimation models extracted from close source software. The data for proprietary source software will be provided, for example, by the widely known ISBSG (International Software Benchmarking Standards Group) data set.

As a result of the previous studies, work will be done on productivity metrics, and its relationship with cost/effort (and value) estimation models, looking for an easy-to-handle estimation model (similar to COCOMO and other modern models and techniques) for libre software development. It should be noted that COCOMO-like models provide the "substitution cost" i.e. the cost of producing a given libre software package if it was to be rewritten from scratch within a firm. "Real cost" productivity is something else: what is the effort that was actually put into writing the libre software package, and how much can that be valued (i.e. opportunity cost of time, based on expected income). For this, we propose to apply a methodology we are already using for estimating that effort based in data we already have available: combining data from past surveys of FLOSS developers (including answers about the effort they have devoted to specific projects during a certain period) and database of software version control data (e.g. CVS). In short, the data from CVS is used to estimate the amount of code produced (or changed) during a certain period by specific individuals, and that information can be matched against the time effort declarations by those individuals in the surveys. The output of this study is a function mapping effort (in person-months) to the amount of produced (or changed) software code. We then apply this mapping function in reverse to a certain percentage of developed FLOSS software (by analyzing the corresponding CVS data and estimating the code output. This way, the total effort put into its development in terms of time can be derived. In addition, the team size can be estimated, as well as the opportunity cost in monetary terms by computing the equivalent salary cost of the estimated time input of the developers. This can then be extrapolated for the total FLOSS code base published, controlling for differences between the sample studied in detail and the entire universe of the code base. The result is a good estimate of the value in effort of the entire primary production of FLOSS software, as well as the equivalent in monetary (opportunity cost) terms for the value of the primary production of FLOSS software. The details of this

tested methodology are based on the work carried out previously at URJC and MERIT. However, we propose to apply this to a much larger range of projects and a much larger data sample than previously achieved.

In general, for this track, a large collection of projects will be considered and analysed. The detailed list will be decided in WP1, but it will for sure include some projects which are interesting, large and well known, such as GNOME. KDE, Apache, OpenOffice, Mozilla, etc., and others of special interest in the European framework (such as Objectweb).

Track 2: Dissemination

The dissemination track is composed of one package, WP7, which will carry, ongoing throughout the project duration, the dissemination of the results of the project. This will be achieved on one hand using the project website, which will publish the results, both reports and software, being produced by the other work packages as soon as they are available. In addition, three international workshops and several (a minimum of four) regional workshops will be held to further increase the dissemination of the results. The international workshops organized by FLOSSMETRICS will be:

- The first workshop will be colocated with a conference with high affluence of libre software developers (such a FOSDEM in Belgium, LinuxTag in Germany or LSM in France). It will be organized by URJC, and will be devoted to explain the plans of the project with respect to the data to retrieve and the expected outcomes and studies. The main aim will be to get feedback of actual developers, which can be integrated in later stages of the project.
- The second workshop will be organized by UM, probably in combination with some event which attracts to business and industrial companies with interest in libre software development. The main aim will be to present the first results, and get feedback about their interest and relevance for the industry, with special attention to the reactions of SMEs and secondary sector companies.
- The third workshop will be celebrated near the end of the project, and its main objective will be to disseminate the results of the project. It will be organized by URJC

UM will be in charge of coordinating and supervising all the workshops, and organizing the received feedback to use it in the project, when appropriate. All the workshps will take place during one day. The entire workshops will be recorded and documented by selected rapporteurs. The workshops will be actively promoted worldwide through the use of the established extensive network of OSS community available through previous projects (including CALIBRE and FLOSSWorld). Promotional material will be developed to reflect the importance of this workshop and to encourage the attendance of relevant stakeholders.

Track 3: Exploitation and validation

This track will deal with validating the usability of the results of the track on empirical foundation and focused studies (datasets, studies and tools), and obtain feedback to later design exploitation plans for them. The validation and feedback retrieval will be performed in two different contexts: SMEs developing or using libre software, or even interested in it (WP8); and industrial players developing libre software and the libre software community at large (WP9) . Based on the feedback obtained in these two contexts, WP10 will be devoted to designing a complete exploitation strategy for the whole libre software industry.

The main milestone of this track will be the delivery of the final exploitation plan, which will comprise the expertise and recommendations collected in the three workpackages, and is one of the main outcomes of the project.

This track is where the industrial involvement in the project is focused. The role of those industrial partners consists in providing high-level inputs and using and disseminating the results. While the intensive effort in the project is performed by the academic partners, the careful selection of industrial partners emphasises the multiplier effect of this industrial contribution.

ZEA is an association of about 20 companies from 10 countries, representing many different commercial and business contexts, which will constitute a good network for both dissemination and feedback collection. In PMS the project counts with Frank van der Liden, who is currently chairing the CALIBRATION forum, and is involved in many other industrial forums with relationship with libre software. This way, the consortium expects also to involve some of those forums, and specifically CALIBRATION[7]. Finally, Conecta is specialized in libre software services, with a good record of consulting in those services, contribution to several international projects, and experience in the needs of SMEs interested in libre software.

This expertise and contacts of the industrial partners will be used to obtain the needed involvement from industry. The rest of the consortium will provide support to them, so that they can actually focus on the tasks in which they can be more productive.

Track 4: Management activities

The management of the FLOSSMETRICS project follows the description of project management in section 6.1. It takes place in the workpackage relating to management, WP12. The management workpackage will be responsible for the three deliverables related to management:

1. A web-based monitoring system providing an up-to-date overview of the status of the various activities of the project including the status of the various project deliverables and workpackages. This is an ongoing deliverable and will start early in the project timeline.
2. An interim management / status report, highlighting the project results and providing the status of the project, and identifying any problems faced and how they were or are planned to be resolved. This will be delivered halfway through the project.
3. A final management / status report, highlighting the final project results including results of the final workshop, and providing the status of the project, and identifying any problems faced and how they were addressed. This will be delivered at the end of the project.

The management workpackage will also monitor risks and problems and plan for their resolution, following the project management structure outlined in section 6.1. As this project is based on measurement and analysis of publicly available data, based on tools and techniques that have already been pioneered by the consortium partners, it faces few specific, significant risk factors.

[7]CALIBRATION was initiated by the EU FP6 Coordination Action Calibre (in which URJC and UM are also partners), with the aim of ensuring that libre software delivers to its true potential for the European secondary software sector, and has presently 14 members in its board, representing both large and small companies (more information is available in http://www.calibre.ie/forum/)

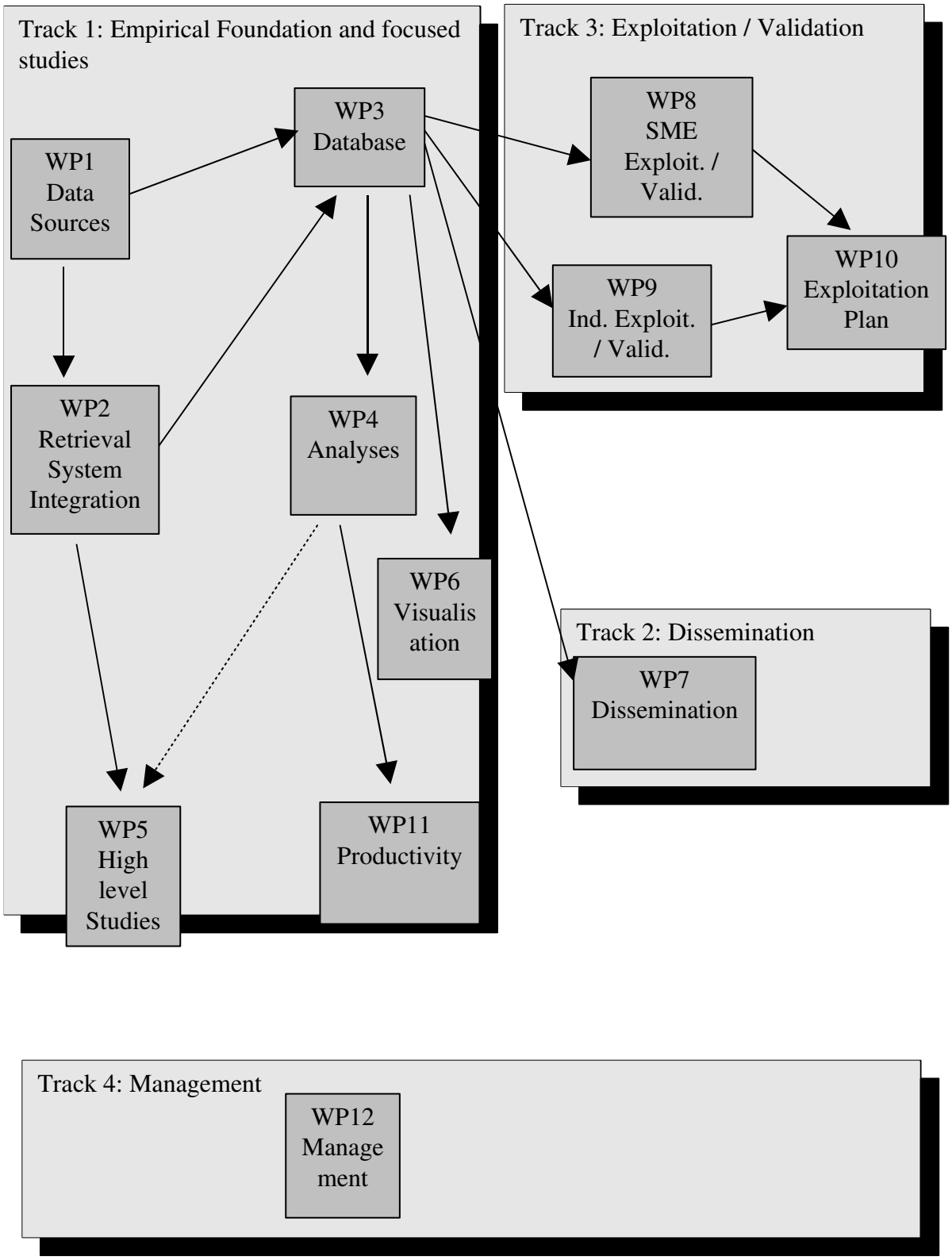
Most of the tools to be used have already been tested by the partners, and they have expertise in their use, in storing the data in large databases, in analysing the data, and in general, in other aspects of the project. The project is also confident of finding enough libre software projects with public repositories to fulfil the goals stated in this Description of Work. In particular, the project has already identified about 30,000 projects with an active CVS repository in SourceForge, and several thousands more in Berlios, Savannah, and others. Bug tracking repositories and mailing lists archives can also be found, by the thousands, in the same hosting sites. In addition, large projects (such as GNOME, KDE, Apache, Mozilla, etc.) maintain also this information for many of their "subprojects".

Therefore, the only risk in this direction seems to be that the project is not able of finding enough repositories with meaningful information of a suitable quality. In this case, heuristics would be used to filter the information obtained from repositories, and increasing the signal to noise ratio (e.g., by removing commits or bug reports, which correspond to automatic activities).

Despite of this, in case the project would not find enough projects with suitable information in the hosting sites (which is the main strategy of search devised), more complete searches will be performed, using catalogs of libre software projects, such as Freshmeat.net, and brute force searches (such as using generic search engines to identify other projects).

7.2. Work planning and timetable

7.3. Graphical presentation of work packages



The diagram above shows the main relationships between the different workpackages, and how they are grouped in tasks. Within Task 1, it can be seen how after identifying the data sources with quantitative information about libre software development (WP1), the work can start in the definition and building of the database (WP3) and in the integration of the different tools that will compose the retrieval system (WP2). Once the retrieval system is complete, the database will be fed, and the work on WP4 (analyses of the retrieved data) and WP6 (visualization of the retrieved data) will take the lead. Finally, once the first analysis are ready, and using again the retrieval system for getting more specific data, high level studies (WP5) and productivity studies (WP11) will be performed.

Within Task 2, dissemination of the main results of the project will be performed (WP7). The first results to disseminate will be the data set stored in the database, to which analysis, studies and exploitation / validation results (Task 2) will follow.

In Task 3, as soon as the information starts to enter the database, but specially when analyses and studies become available, the exploitation and validation of those outputs for the specific needs of SMEs (WP8) and software industry in general (WP9) will lead to the design of an exploitation plan (WP10).

Task 4 will provided the needed management support to the rest of the activties in the project (WP12).

7.4. Workpackage List

Work-package[8]	Workpackage title	Deliverable No[9]
WP1	Data Sources	**1.1, 1.2, 1.3**
WP2	Retrieval system integration	**2.1, 2.2**
WP3	Database	**3.1, 3.2**
WP4	Analyses	**4.1, 4.2**
WP5	High level studies	**5.1, 5.2**
WP6	Visualisation	**6.1**
WP7	Dissemination	**7.1.1-6, 7.2.1-10, 7.3.1-3, 7.4, 7.5, 7.6**
WP8	SME Exploit./Valid.	**8.1.1-3**
WP9	Ind. Exploit./Valid.	**9.1,9.2**
WP10	Exploitation plan	**10.1**
WP11	Productivity	**11.1, 11.2, 11.3**
WP12	Management	**12.1, 12.2.1-10, 12.3.1-3, 12.4.1-3, 12.5.1-3, 12.6, 12.7, 12.8, 12.9**
	TOTAL	

[8] Workpackage number: WP 1 – WP n.
[9] Deliverable number: Number for the deliverable(s)/result(s) mentioned in the workpackage: Dx.y

7.5. Deliverables List

Deliverable No	Deliverable title	Delivery date	Nature[10]	Dissemination level[11]
D1.1	Study of Available Tools	4	**R**	**PU**
D1.2	List of Selected Projects	7	**R**	**PU**
D1.3	Repository Finder	9	**P**	**Pu**
D2.1	Design of Retrieval System	4	**R**	**PU**
D2.2	Implementation of the Retrieval System	12	**P**	**PU**
D3.1	Database Specification	8	**R**	**PU**
D3.2	Database	13	**P**	**PU**
D4.1	Classification Report	15	**R**	**PU**
D4.2	Updated Database V1	18	**P**	**PU**
D5.1	Software Engineering Studies	25	**R**	**PU**
D5.2	Updated Database V2	25	**P**	**PU**
D6.1	Visualisation Prototype	23	**P**	**PU**
D7.1.n	Website (2 updates per year)	1 / 6 /13 / 18 / 24 / 29	**O**	**PU**
D7.2.n	Workshops (1-3)	9 / 18 / 29	**O**	**PU**
D7.3.n	Results and impact (1-3)	12 / 24 / 30	**R**	**PU**
D7.4	Project presentation	1	**R**	**PU**
D7.5	Raising public participation and awareness	30	**R**	**PU**
D7.6	Publishable Summary Final Report	30	**R**	**PU**
D8.1.n	Guide for SMEs (1-3)	12 / 19 / 26	**R**	**PU**
D9.1	Industrial inputs to exploitation	18	**R**	**PU**
D9.2	Industrial validation	29	**R**	**PU**
D10	Exploitation Plan	29	**R**	**PU**
D11.1	Metrics dictionary	26	**R**	**PU**
D11.2	Productivity study report	26	**R**	**PU**
D11.3	Cost/effort estimation study	29	**R**	**PU**
D12.1	Monitoring System	30	**P**	**PU**
D12.2.n	Progress report (1-10)	3 / 6 / 9 / 12 /15 /18 / 21 / 24 / 27 / 30	**R**	**CO**
D12.3.n	Activity report (1-3)	12 / 24 /30	**R**	**CO**
D12.4.n	Management report	12 / 24 /30	**R**	**CO**
D125.n	Financial distribution report	12 / 24 /30	**R**	**CO**
D12.6	Final activity report	30	**R**	**CO**
D12.7	Final management report	30	**R**	**CO**
D12.8	Final financial distribution report	30	**R**	**CO**
D12.9	Final plan for using and disseminating knowledge	30	**R**	**PU**

[10] Please indicate the nature of the deliverable using one of the following codes: R = Report; P = Prototype; D = Demonstrator; O = Other

[11] Please indicate the dissemination level using one of the following codes:

PU = Public

PP = Restricted to other programme participants (including the Commission Services).

RE = Restricted to a group specified by the consortium (including the Commission Services).

CO = Confidential, only for members of the consortium (including the Commission Services).

Deliverable No	Deliverable brief description
D1.1	Study of tools available for the data retrieval (downloading and raw analysis of artifacts in repositories related to libre software development)
D1.2	Complete list of projects for which information will be retrieved, with detailed information about their repositories
D1.3	Tool to find repositories with information about libre software development
D2.1	Design of the system for data retrieval
D2.2	Implementation of the system for data retrieval
D3.1	Specification of the database which will store all the data retrieved from repositories
D3.2	Implementation (and data feed) of the database (prototype version)
D4.1	Report on the classification of libre software projects, according to the characteristics of the data obtained about their development
D4.2	Implementation (and data feed) of the first revision of the database, having into account the lessons learned from the first analyses
D5.1	Studies (from an empirical software engineering point of view) of the projects corresponding with the data sets in the database
D5.2	Implementation (and data feed) of the second revision of the database, having into account the lessons learned from the first studies
D6.1	Prototype of visualisation system for the data sets in the database, allowing for visualization of several different aspects of the data
D7.1.n	Website of the project (2 updates per year)
D7.2.n	Three workshops for the dissemination of the results of the project, and to get feedback from the target communities (SMEs, companies with interest in libre software, libre software developers)
D7.3.n	Three reports on the results and impact of the project (each for an annual period)
D7.4	Documents presenting the project
D7.5	Report on activities performed for raising public participation and awareness
D7.6	Summary of the final report, written to be useful for dissemination
D8.1.n	Three versions of the “Guide for SMEs”, a report which will build on the studies in the project to address the needs of SMEs with respect to libre software development
D9.1	Report on the industrial inputs for the exploitation of the results of the project
D9.2	Validation by software industry (including secondary sector) interested in libre software of the outcomes of the project
D10	Plan for the exploitation of the results of the project
D11.1	Dictionary of metrics related to activity and effort estimation
D11.2	Study on the productivity of libre software projects
D11.3	Study on the estimation of activity, effort and cost, suitable for libre software development
D12.1	Web-based system for monitoring continuosly the progress of the project
D12.2.n	Quartely reports (10) on the progress of the project
D12.3.n	Annual reports (3) on the activities performed by the project
D12.4.n	Annual reports (3) on the progress of the project from a management point of view
D12.5.n	Annual reports (3) on the distribution of financial resources
D12.6	Final report on the activities performed by the project
D12.7	Final report on the progress of the project from a management point of view
D12.8	Final report on the distribution of financial resources
D12.9	Final report on how to use and disseminate the knowledge produced by the project

7.6. Work package Descriptions

Workpackage number	**1**	**Start date or starting event:**	**Month 1**
Workpackage name	**Data Sources**		

Objectives (Data Sources)
Identify, analyse and evaluate all possible data sources for quantitative data on FLOSS projects, with special emphasis on data publicly available on the Net.

Description of work
The data sources of information related to libre software development are complex and many. They have to be carefully identified in kind (types of repositories, types of systems used to maintain those repositories) and in element (different repositories spread in the Net). This workpackage will address this identification, leading in the end to the decision of which projects will be analysed. Of course, in many cases the decision is already made: FLOSSMETRICS will analyze well known projects such as GNOME, KDE, Apache, Mozilla, or ObjectWeb. However, specially in the case of small projects, a careful analysis and identification is needed. This work package includes the following tasks:
Task 1: Identification of possible sources of quantitative data on FLOSS projects. These include primary sources: systems directly used in the development process (like source code repositories, bug-tracking systems, mailing lists, news forums, package repositories, project development web sites etc.), and secondary sources, which offer information not directly produced by the projects themselves, (such as web search engines, libre software portals, etc). This task will include also identification of different engineering, socio-economic, and organisational questions that can be addressed by such information, and an assessment of the reliability and usability of each data source.
Task 2: Survey of tools used in libre software development. Several tools are widely used within the developer community to manage various aspects of the development process including coordination, organisation, feedback, version control, bug tracking. These tools result in data and metadata that can be excellent sources of analysis. This task will result in a list of specific tools, and an estimate of their usage in the community. Examples include CVS, Subversion, Bugzilla etc.
Task 3: Analysis of those tools on which information is contained and identification of how data resulting from their usage by libre software developers can be accessed and processed automatically. Analysis of underlying data structures of these tools to uncover overlaps and/or possibilities for common extraction mechanisms.
Task 4: Decision on which FLOSS projects to use for prototypical in-depth data extraction and analysis (in the following work packages). This choice is limited by the tools and systems used by each project to store information which will be collected (source control system, mailing lists, bug tracking systems, etc). At this point the project will also check the total number of identified projects, to ensure that a reasonably large quantity is included in the study (according to the stated goals of the project).

Deliverables
D1.1 Study of available tools. This is a report with an exhaustive study of the list of tools used by libre software projects in the development process, and the traces of information each leaves for later analysis. Includes a survey of their use across projects (qualitative and quantitative) and their suitability as accessible, usable sources of information.
D1.2 List of selected projects. Report on the list of projects selected, the kind of data available for each, and the exact location of information repositories for all of them.
D1.3 Repository finder. Tool that tries to automatically find the usual repositories of information related to development tools for a given project (using heuristics, like if the project is in Sourceforge, the CVS would be in cvs.sourceforge.net)

Milestones and expected result
M1.1: Completion of deliverables D1.1, D1.2 and D1.3
Expected results: Complete definition of the sources of data for the rest of the project, including some tools that help to automate the future identification of new sources (Repository finder).

Workpackage number	**2**	**Start date or starting event:**	**Month 1**
Workpackage name	**Retrieval system integration**		

Objectives (Retrieval system integration)
Building a system (by integrating already available tools into an integrated platform) capable of automatically fetching and analysing data related to libre software projects. The system will have a modular architecture, to simplify the reuse of already existing tools.

Description of work
This work package will integrate the fetching system (which will perform also some preliminary, automatic analysis of fetched data), and will feed the database system (see WP3). The developed system will reuse as much existing modules as possible (for instance, modules calculating SLOC counts, authorship data, quality metrics, etc. do already exist, made by the partners of the project or available as libre software from external sources).
This work package includes the following tasks:
Task 1: Extensive review of the tools already available as libre software for the retrieval and analysis of information related to libre software projects, and study of how they could be integrated together.
Task 2: Integrating the identified tools in a system capable of controlling the process of automatically finding, retrieving and storing raw data, and injecting it into the database.
It is important to notice in this workpackage that most of the tools to retrieve information from repositories have been built and/or are already in extensive use by the partners of the consortium. The workpackage will basically lead with combining and streamlining them so that the massive retrieval of data can be done automatically and in a reliable way.

Deliverables
D2.1 Design of retrieval system. Will include a study of the tools needed and available to retrieve information of each specific repository; the specification and design of the system; and the integration architecture.
D2.2 Implementation of retrieval system. Includes testing with some repositories.

Milestones and expected result
M2.1: Completion of deliverables D2.1 and D2.2
Expected results: integrated system capable of fetching data about a project, and doing a preliminary, automated analysis of it.

Workpackage number	**3**	**Start date or starting event:**	**Month 5**
Workpackage name	**Database**		

Objectives (Database)
Putting together an extensive database, available via Internet, with raw and processed data about libre software projects.

Description of work
All the data related to libre software projects being analysed will be stored and managed in the database system. The version built in this work package will have the basic functionality of storing all the fetched data and the results of the analyses performed on them. Several interfaces will be available for retrieval of data: web interfaces for querying about specific data, more general project reports (also available via web), and complete raw data corresponding to a specific project, as an XML file (useful mainly for sharing raw data).
This work package includes the following tasks:
Task 1: Study on the hardware and software requirements for the database management system and its interfaces (including a web-based one).
Task 2: Specification and design of the interfaces to the database systems
Task 3: Development of the first version of a test bed of the database and some of its interfaces, and testing using real projects
Task 4: Development of the final version of the database and its interfaces, and putting the system into production (routinely searching for projects, retrieving its data, adding it to the general database, updating of data periodically, etc.)
The database will be fed with data for thousands (maybe tens of thousands) of projects, by retrieving the information available in the repositories identified in WP1, using the retrieval system integrated in WP2.

Deliverables
D3.1 Database specification. Including hardware requirements and data model.
D3.2 Database. Final version of this step of the database (see other workpackages below for other steps).

Milestones and expected result
M3.1: Completion of Deliverables D3.1 and D3.2
Expected results: comprehensive database with development information about libre software projects. Although the database system will be completed as deliverable D5.2, the system will continue fetching and updating data automatically until the end of the project.

Workpackage number	**4**	**Start date or starting event:**	**Month 9**
Workpackage name	**Analyses**		

Objectives (Analyses)
Developing a scheme for classification of projects, after careful analysis and correlation of the fetched processed data.

Description of work
The fetched processed data, after basic analysis, will be correlated and studied statistically, looking for parameters useful for classifying libre software projects. Several classification schemas will be provided, each one offering useful data from a different point of view (development rate, volume, efficiency, quality, etc.)
This work package includes the following tasks:
Task 1: Identification of relationship between variables determined and extracted in WP 2 and 3.
Task 2: Identification of possible discriminatory criteria computed from existing variables (e.g. commits per programmer per month or efficiency scores from applying data envelopment analysis) and their relationship with other variables. Computation of these criteria and storage in the data base.
Task 3: Using the results from task 1 and 2, development of a classification scheme for FLOSS projects.
Task 4: Application of the classification scheme for the projects in the data base and storage of the results in the data base

Deliverables
D4.1 Classification report. Report detailing the results of the application of analytical techniques and classification
D4.2 Update Database V1. Update of the database system to allow storage of newly developed criteria and classification results.

Workpackage number	**5**	**Start date or starting event:**	**Month 14**
Workpackage name	**High level studies**		

Objectives (High level studies)
Demonstration of possibilities for focussed studies and the development of prediction models using the results from the data base. Consolidating findings from several FLOSS aspects into a coherent picture.

Description of work
With the data obtained in WP3 and the classifications of WP4, more in-depth analysis will be performed, from a more classical software engineering perspective. These studies will show the usefulness of the obtained data for advancing the knowledge base on libre software engineering, at least in the fields of productivity, effort and quality prediction models.
This work package includes the following tasks:
Task 1: Decision on which aspects of FLOSS development to study in depth, based on data available in the data base from WP 3 and WP4.
Task 2: Performance of focussed studies and development of prediction models (e.g. for productivity, effort or quality) for selected projects from the data base and storage of results in the data base.
Task 3: Consolidating finding from different aspects like software engineering and economics to form a coherent picture of FLOSS projects.
Task 4: First steps towards the application of traditional prediction models to FLOSS projects overall (not only single aspects), e.g. using techniques like dynamic system modelling.

Deliverables
D5.1 Software engineering studies. Report detailing the results from the focussed studies and possible further studies.
D5.2 Update Database V2

Milestones and expected result
M5.1: Completion of deliverables D5.1 and D5.2
Expected results: Better understanding of studied libre software projects from a software engineering point of view (best practices, relationship with development models, with prediction models, etc.)

Workpackage number	**6**	**Start date or starting event:**	**Month 9**
Workpackage name	**Visualisation**		

Objectives (Visualisation)
To develop a prototype (using already available tools) which allows for visualisation of the results in the database, both produced from data extraction (WP 3) and application of analytical (WP 4) and highe level studies (WP 5).

Description of work
The huge quantity of data available in the database system would be much more useful with tools which help to visualise and understand them, and what is more important, their relationships and implications. This work package will focus on several visualizations of the stored data, each aimed to shed some light on a class of problems interested for the research in this area. For the visualization, already available libre software tools, such as OpenDX, http://opendx.org, will be used.
This work package includes the following tasks:
Task 1: Analysis of potential user groups (researchers, FLOSS participants, public,...) and their information needs.
Task 2: Decision on the information from the database to be presented visually, in accordance with findings from task 1.
Task 3: Analysis of possible forms of information visualisation especially for large data sets, including fish-eye views or similar approaches from the human-computer-interaction field.
Task 4: Implementation of a prototype to support visualisation of chosen information with appropriate techniques, using already available tools.
Task 5: Application of prototype from task 4 to available data and publication on the web site.

Deliverables
D6.1 Visualisation prototype

Milestones and expected result
M6.1: Completion of deliverable D6.1.
Expected results: Visualisation techniques for helping in the reasoning about libre software development.

Workpackage number	**7**	**Start date or starting event:**	**Month 4**
Workpackage name	**Dissemenation**		

Objectives (Dissemination)

To disseminate the results of the project (basically, a deeper understanding of specific issues related to the role of libre software in the Strategic Objectives), through interaction with the target community, as widely as possible.

Description of work

All FLOSSMETRICS deliverables are intended for public release as soon as possible (while addressing privacy and practical considerations). A target community comprising three groups has been determined:

- the academic community (including statistical departments of government and multilateral bodies as well as academic and research institutes/universities), which will be interested in the empirical results, analytical framework and deeper understanding of the open source phenomena, as well as the methodological issues highlighted by the study.
- the business community, which will be interested in identifying models whereby they can generate commercial value out of open source as an efficient development environment (especially important to SMEs) as well as the insights FLOSSMETRICS will provide in terms of understanding libre software development as a model of collaborative production.
- the libre software developer community itself, greatly appreciate more and specific understanding about its functioning and methods as seen by the enormous community interest in the results of the FLOSS project. The developer community benefit from FLOSSMETRICS findings on their own management, structure and project inter-dependency attributes, which are otherwise not accessible. This developer community will be the primary target group for FLOSSMETRICS.

The dissemination workpackage will involve the maintenance of a public website providing access to all tools, databases and research results from the FLOSSMETRICS project on an on-going basis.

It will also include the organisation a three workshops, at the beginning, middle and end of the project to involve researchers and industry in active participation and discussion of the project results and especially plans for further development and exploitation. SMEs not directly involved in libre software development (but interested in it) will also be invited, to get their feedback and ensure that their needs are covered by the project.

In fact, the dissemination of the results at workshops co-located with libre software conferences will also ensure that the results are available and known to the libre software community in general, including those volunteer developers so fundamental.

In general, all workshops will be designed so that they will also be of interest for SMEs, and effort will be made to facilitate their attendance as much as possible. In this line, the FLOSSMETRICS website will also include a section with information specific for SMEs with interest in libre software.

This workpackage will also include 2.5 person-months of work related to concertation activities (see detail in section 9.3). Part of those activities are already contemplated in the rest of the workpackage, but others are quite specific, and will depend on the requirements by the Commission and other institutions.

Deliverables

D7.1.n (1-6) Website. Will host increasingly more information as the project proceeds (two deliverables per year)

D7.2.n Workshop (1-3)

D7.3.n Results and impacts (1-3), one per year

D7.4 Project presentation

D7.5 Report on raising public participation and awareness

D7.6 Publishable summary final report

Workpackage number	**8**	**Start date or starting event:**	**Month 8**
Workpackage name	**SME Exploit./Valid.**		

Objectives (SME Exploitation/Validation)
To validate the usability of, and develop an exploitation plan for, the project's research results and tools within an industrial open source development environment from the SME perspective.

Description of work
Many SMEs exploiting open source software provide the so called "integration and installation" services, consisting of three separate phases: identify the customer's needs; find the suitable package, or set of packages, that can solve the problem identified; provide the installation, training, and post-install support. A particularly weak point at the moment is the second phase, i.e., it is extremely difficult to find, in the many packages available, which one is the most appropriate, from a long term perspective. It is in fact important to understand the "vitality" of a project, to provide at least reasonable expectations about the continued development of a chosen package; this must be done to the maximum extent in an automated way. Partner 4 (Conecta) will test the applicability of the FLOSSMETRICS methodology and results against a set of software packages with which Conecta has had previous experience, in order to evaluate the usefulness of the data in the package selection phase, and will prepare a short report targeted at SMEs on how to use the results from FLOSSMETRICS to improve the quality and speed of the package selection process.
The results, along with potential business and exploitation benefits for SMEs, will be summarized in a "FLOSSMETRICS guide for SMEs", after feedback and consultation with external SMEs. This guide will be targeted not only to those SMEs already using libre software, but also to those which could be interested in doing so, by lowering the entry barriers and uncertainty which prevent them to embrace libre software as solutions for their IT problems.
One of the specific ways to address this goal will be to include a chapter in business models related to libre software for SMEs, partially based in a deliverable written for the COSPA project (in which Conecta has participated), extending its results to include the quantitative evidence provided by FLOSSMETRICS. The guide will also include specific recommendations for SMEs about how to compare libre software products based on the vitality of their communities and development projects, including sets of parameters (obtained from the extensive study) that may characterized "mean" cases (so that companies can decide that a given project is performing over, or below, the mean, for instance). It will also include information on which aspects of the development process of a given project could be of interest for an SME considering using it in the context of its business plan.
This guide will be distributed through the network of clients and related SMEs of both Conecta and ZEA, to give a wide audience, and being able of getting also their early feedback, to validate it, and to help to better address the needs of such companies (with specific attention to detect whether the needs of those companies not currently involved in libre software are addressed, and take corrective measures if that is not the case).
The guide will also be presented at some conference with relevant attendance of SMEs interested in the use of libre software.
There will be three versions of the guide, reflecting different stages of feedback and results of the project. Therefore, the final version of the guide will in fact be the result of an incremental process. The project will do its best to publish the final version of the guide in book-quality, under a Creative Commons (or similar) license, if possible.

Deliverables
D8.1.n: Guide for SMEs (one per year)

Milestones and expected result
M8.1: Completion of deliverable D8.1.3
Expected results: Understanding of the implications of the research results on SMEs

Workpackage number	**9**	**Start date or starting event:**	**Month 14**
Workpackage name	**Industrial Exploit./Valid.**		

Objectives (Industrial Exploitation/Validation)
To validate the usability of, and develop an exploitation plan for, the project's research results and tools within an industrial libre software development environment

Description of work
The approach in this workpackage centres on soliciting inputs and feedback from industrial players external to the consortium, and the open source developer community in general. While the open source developer community is expected to respond and provide feedback thanks through the dissemination in WP7, a close collaboration with specific industrial players will be taken up in this workpackage. In particular, PMS has expressed its support for the FLOSSMETRICS project and their interest in collaboration by joining the consortium. This will allow for a much more complete validation and exploitation of the research results. The person-months for this WP are limited to those contributed by consortium members, but other industrial players are expected to contribute their own person-months in validating and using FLOSSMETRICS results based on their individual interests, and providing the FLOSSMETRICS consortium with feedback, through this WP.
Task 1: Solicit on-going feedback from industry intended regarding the work of the R&D workpackages during the development of data extraction tools, the database structure and the sample base of open source projects to be included in the database, in order to ensure that they best meet the needs of open source development in an industrial environment
Task 2: Solicit from industry feedback based on their testing and validation of the data analysis tools and benchmarking models resulting from the R&D work packages on open source projects, in particular, projects involving employees of PMS and other external industry players consulted.
Task 3: Provide inputs to the exploitation plan for industrial users (to be developed by UM in workpackage 10) by identifying ways in which the research results and data analysis tools, benchmarking and predictive models can be usefully applied within an industrial open source development environment, based on the feedback provided from industrial players such as PMS and ZEA. Provide further inputs as to whether such tools may also be used in the monitoring, benchmarking and management of proprietary software development in an industrial environment.
In addition, PMS is experimenting with inner source development. This means that open source practices are performed within the company borders. Within this work package the effectivity of this way of working will be evaluated. The methods and tools developed in the FLOSSMETRICS project will be used, and feedback on these will be used as input for the R&D project. In particular, PMS will perform evaluations on the ongoing development projects; providing on-going feedback on the usability of the methodology and tools, and on the usefulness of the results; providing inputs to the exploitation plan for industrial users by identifying ways in which the research results and data analysis tools, benchmarking and predictive models can be usefully applied within an industrial inner source development environment. providing further inputs as to whether such tools may also be used in the monitoring, benchmarking and management of proprietary software development in an industrial environment.

Deliverables
D9.1 Industrial inputs to exploitation
D9.2 Industrial validation. External Industrial Validation Report.

Milestones and expected result
M9.1: Completion of deliverable D9.2
Expected results: Understanding of the implications of the research results on industry

Workpackage number	**10**	**Start date or starting event:**	**Month 20**
Workpackage name	**Exploitation plan**		

Objectives (Exploitation plan)
Based on inputs from industry and including the SME perspective (WP 8-9), to report on the validation of the project's research results and tools within an industrial open source development environment. Based on industrial and SME inputs (WP 8-9), the development of a comprehensive exploitation strategy on how these research results and tools can best be used and exploited within the open source software industry: by large firms, SMEs and the broader open source developer community.

Description of work
To guarantee a continuation after the end of the project, an exploitation plan will be prepared to identify potential markets and channels for the data and tools that are resulting from the WP 1-6. There will be two potential communities, the commercial software firms, and the companies that are using libre software within their products, that will probably the ideal target for the exploitation activities. The exploitation plan will provide an overview of different, potential business plans based on the data and the tools.
The exploitation plan will include specific recommendations and guidelines about how the outputs and results of the FLOSSMETRICS project could be used by the software industry in general (with specific attention to the secondary sector), and by those industries involved, or planning to be involved, in libre software development (based on the experience of the partners of the consortium, and the outputs of WP 9.
The project will do its best to publish the final version of the exploitation plan (or at least the part of it dealing with recommendations and guidelines) in book-quality, under a Creative Commons (or similar) license, if possible.

Deliverables
D10.1 Exploitation plan

Milestones and expected result
M10.1: Completion of deliverable D10.1.

Workpackage number	**11**	**Start date or starting event:**	**Month 14**
Workpackage name	**Productivity**		

Objectives (Productivity)
Evolve economic metrics based on software engineering metrics, perform specific economic and productivity studies, especially the application of comprehensive cost/productivity estimation models to libre software projects.

Description of work:
In this workpackage, the results from software engineering will be translated to economic, productivity and activity metrics. This will be used to develop a cost estimation study for libre software projects.
Task 1: Build a "metrics dictionary" in order to interpret the metrics identified by software engineering tools in socio-economic terms. This aims to answer questions such as: is a CVS-"committer" the original producer of a piece of software? Its owner? A value-adder (editor) rather than original producer (author)? If source lines of code (SLOC) for a particular project represents *net* production, what is the equivalent to *gross* production?
Task 2: Specific studies on economic and productivity metrics
Task 3: Apply a comprehensive cost estimation model to estimate cost in terms of time, effort and money given characteristics of a libre software project. Based on selected output metrics (such as SLOC, number of developers and developer organisation structure) together with survey data on actual developer time and effort (from the FLOSS developer survey; the FLOSS-US survey which asked developers for time spent on specific named projects; anthropological study of developer communities from the on-going FLOSSPOLS project; and possibly data from additional survey questions added to the developer survey in the on-going FLOSSPOLS project).
FLOSS, FLOSS-US and FLOSSPOLS are projects in which UM participates or has participated, which means that has complete access to them.

Deliverables
D11.1 Metrics dictionary
D11.2 Productivity study report
D11.3 Cost/effort estimation study (for libre software projects)

Milestones and expected result
M11.1: Completion of deliverables D11.1, D11.2 and D11.3.
Expected results: Understanding of the libre software development models from the viewpoint of productivity and cost estimations and models.

Workpackage number	**12**	**Start date or starting event: Month 1**
Workpackage name	**Management**	

Objectives (Management)
The overall management of the consortium and the project, and providing a picture of the on-going status of the project's progress including the reporting of delays or other problems related to workpackages.

Description of work
An on-going priority will be to provide useful and timely feedback to the IST Programme and related EC measures, to support both programme planning and the activities of individual projects and networks especially in connection with the promotion and support of libre software, but also the broader process of supporting e-government and interoperability issues, taking into account of the Commission's own views on what types of input and support are appropriate and of most value.
A monitoring system will also be available on the web-site to indicate the current status of the project and progress of various work-packages so that the timeliness of the project can be maintained and managed efficiently. This workpackage will also provide for meetings of the coordinators and the overall project management and integration efforts (approximately one meeting every six months, which makes a total of about seven meetings).
The project will report to the EC with different periods:

- Every three months, a progress report will be delivered. This report will detail the activities of the reporting period, the status of workpackages and deliverables, the deviations from plans and the correcting measures taken to overcome those problems, the effort devoted by partner to activities of the project during the reporting period, etc.
- Every year, three annual reports will be delivered, on activity (detailing all the activities of the reporting period, and putting them in the context of the evolution of the project, etc); management (detailing meetings during the period, evolution of the effort devoted by partner, main decissions taken, deviations and correcting measures, etc.); and financial distribution between contractors.
- At the end of the project, final reports on the same topics will be delivered. These reports will show the outcomes of the project, and will put into context the information in the annual reports.

The project coordinator will take care of the merging and edition of the information provided by the partners to produce consolidated reports for all the consortium.

Deliverables
D12.1 Monitoring system
D12.2.n Progress report (1-10), every three months
D12.3.n Periodic activity report (1-3), one per year
D12.4.n Periodic management report (1-3), one per year
D12.5.n Periodic report on the financial distribution between contractors (1-3), one per year
D12.6 Final activity report
D12.7 Final management report
D12.8 Final report on the financial distribution between contractors
D12.9 Final plan for using and dissemination knowledge

Milestones and expected result
M12.1: Completion of deliverables D12.3.1, D12.4.1 and D12.5.1
M12.2: Completion of deliverables D12.3.2, D12.4.2 and D12.5.2
M12.3: Completion of deliverables D12.6, D12.7, D12.8 and D12.9
Ongoing during FLOSSMETRICS project duration.

8. Project resources and budget overview

9. Other issues

9.1 Ethical Issues

The data retrieved in this project will always be public data, available in publicly accessible repositories. However, sometimes this data can contain information, which can easily be tracked to a given person (or to a mail address, for that matter). Although it would be really rare that private information could be accessed this way, special measures are taken to ensure that the developers about which data could be retrieved are protected from any possible harm.

For this matter, participants are aware of possible ethical problems and will work in full compliance with directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. They will also follow the national legislations of the countries in which they carry out their activities.

However, neither European nor national directives inform on how one should undertake studies on publicly available data which, when crossed or correlated could lead to unexpected, and potentially problematic, results, or may become sensitive in the context in which it is collected and presented. To avoid possible problems in this respect, the data will be offered as dissociated as possible from personal data, and when that data has to be stored (for instance, to ensure unicity of identities), it will be encoded appropriately. A logical firewall will be designed as a part of the database implementation so that the personal identities of developers are hidden even to the researchers themselves, except for a very small group which will need it for the aforementioned unicity checks. In most cases, an encoding algorithm will be used as a part of the design of this logical firewall (so that nobody without the encoding algorith can learn about the personal identities), but in some cases (such as identification of different email addresses for the same person) access to the actual data is needed.

Ethical issues form

A. Proposers are requested to fill in the following table

Does your proposed research raise sensitive ethical questions related to:	**YES**	**NO**
• Human beings		NO
• Human biological samples		NO
• Personal data (whether identified by name or not)	YES	
• Genetic information		NO
• Animals		NO

9.2 Gender issues

Gender and libre software engineering

It is a well-known fact that in the field of information and communication technologies there is a divide in the labour force: whereas women are almost equivalently represented in areas in which only low qualification is necessary, such as data entry, they are under-represented in

high-skilled jobs such as programming or system administration. Current research shows a share of women of between 20% and 25% in this high-qualified sector. These figures are in line with graduation figures in computer science. Literature trying to explain the low ratio of women in high-skilled computer jobs in general refers to arguments such as low self esteem, discouragement by the social environment during the early phase of personal development, lack of female role models, and a generally male image of computing in public media[12].

In the field of libre software development however, the figures are even lower. All large-scale surveys on programmers indicate that these communities are extraordinarily male dominated. Whereas surveys vary as to other demographic data such as age or nationality of the participants, the gender question is unambiguous: The WIDI, FLOSS and BCG/OSDN survey all show that less than 2 % of all open source/free software developers are female[13]. While there may be a self-selection bias in these surveys, it is unlikely that more than 5% of developers are women. Notwithstanding these clear findings, current research undertaken on the phenomenon generally does not address this problem at all, and normally presents the libre software phenomenon in a gender-neutral manner. At present there is no systematic study explaining this large gender gap between proprietary software development on the one hand and open source/free software development on the other. This is one of the core topics of the on-going FLOSSPOLS project, which takes an ethnographic approach to addressing this issue.

FLOSSMETRICS being an SSA involving almost entirely data retrieval and studies cannot really address these gender issues. However, we have given thought to this and there is one partial solution to including gender as one of the attributes gathered with the other technical and socio-economic metrics. Many of the tools developed in WP2 will extract names of developers from various sources, such as CVS version control or mailing list data or even source code itself. Among the post-processing methods used will be one that looks up the developer name in standard dictionaries of names to identify the gender, if possible. This can then be stored as an attribute of the developer, and used as a metric in all other analyses. It remains to be seen whether any statistically significant or useful results will arise from this approach, but at least it addresses the gender dimension for the first time, since this technique has never before been applied in open source metrics studies.

9.3 Concertation activities

The project will actively participate in the activities organised at programme level relating to the IST area with the objective of providing input towards common activities and receiving feedback (e.g. from clusters), contributing to area and portfolio analysis, offering advice and guidance and receiving information relating to IST programme implementation, standards, policy and regulatory activities, national or international initiatives, etc.

The project participants will also commit themselves to support the organisation of an annual conference by providing papers, participating in technical programme committee, chairing sessions, reporting, etc.

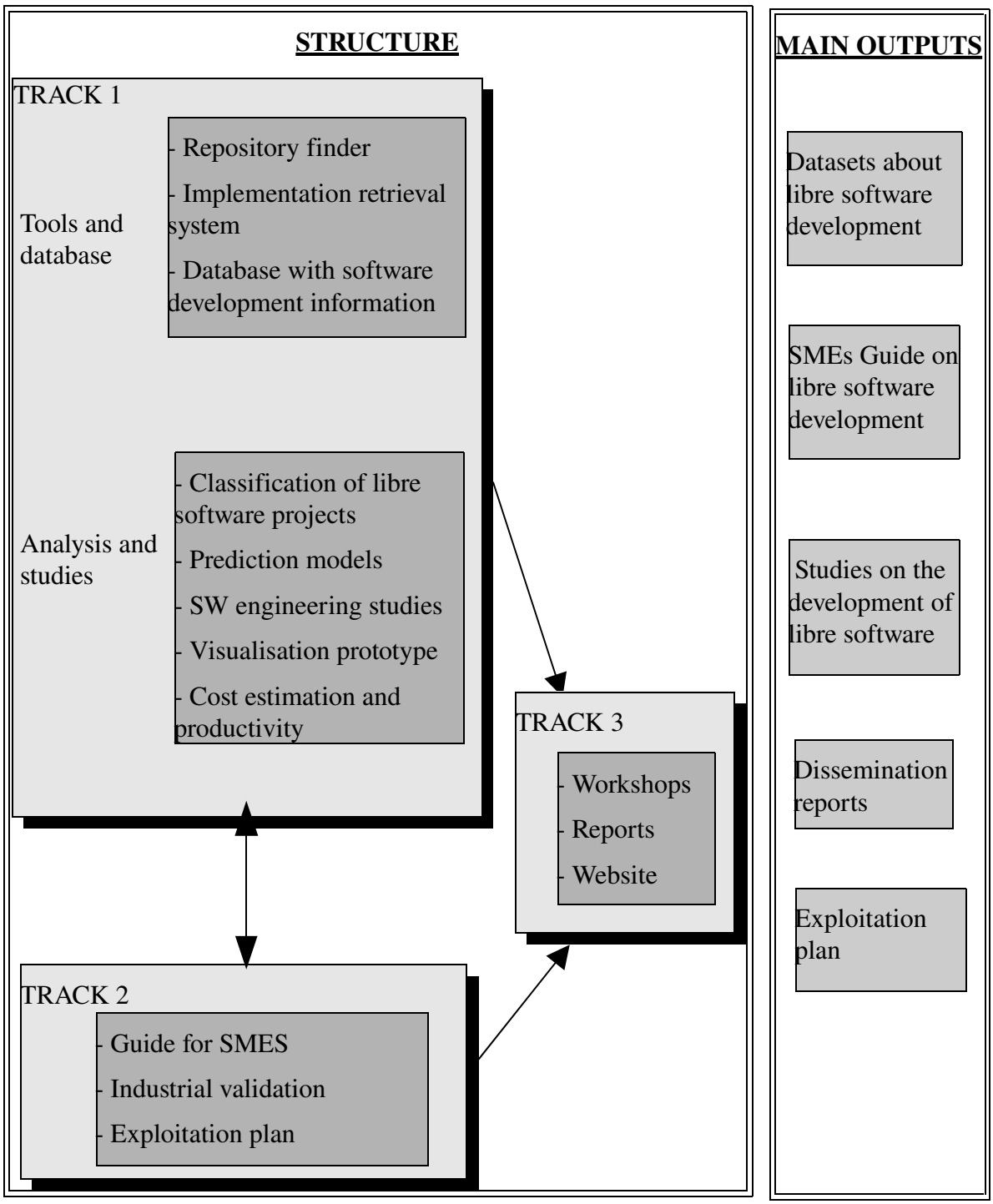
The project participants will help in developing dissemination material that can be used for communication towards the general public. For instance, by developing a video demonstrating the results of the project, by having articles about the project in local newspapers, featuring

[12] Furger 1998; Fisher, Margolis, and Miller 1997; Spertus 1991.
[13] Robles 2001; BCG 2002; FLOSS 2002

the benefits of the research carried out for the community reading the newspaper, or contributing to the development of public relations and state-of-the-art brochures.

9.4 Roadmap

The main objective of FLOSSMETRICS is to construct, publish and analyse a large scale database with information and metrics about libre software development coming from several thousands of software projects, using existing methodologies, and tools already developed. The project will also provide a public platform for validation and industrial exploitation of results.



The activities of the project are organized in four tracks.

- Track 1 is the core of the project. In its context, the largest database in the world with quantitative information about libre software development will be built. The datasets in it, which will be made public throughout the Internet, are expected to develop into a resurce widely used by the research community. The project will also analyse and study in detail those datasets, from several points of view (ranging from empirical software engineering to activity and effort estimation). Some visualization techniques will also be applied to them, to better understand the millions of records stored in the database.
- Track 2 will be devoted to the dissemination of the results of the project. Achieving a large impact will depend heavily on the appropriate dissemination of results. Several communities will be the focused targets for this dissemination: SMEs with interest in libre software development, the software industry in general (and in particular, those companies considering libre software development models), and libre software developers themselves.
- Track 3 will put the results of the project in the context of the target communities. In it, feedback from SMEs and the software industry will be used to validate, improve and focus the main outcomes of the project. In addition, specific guides and recommendations will be written, summarizing the results of the project, and applying to the target domains.

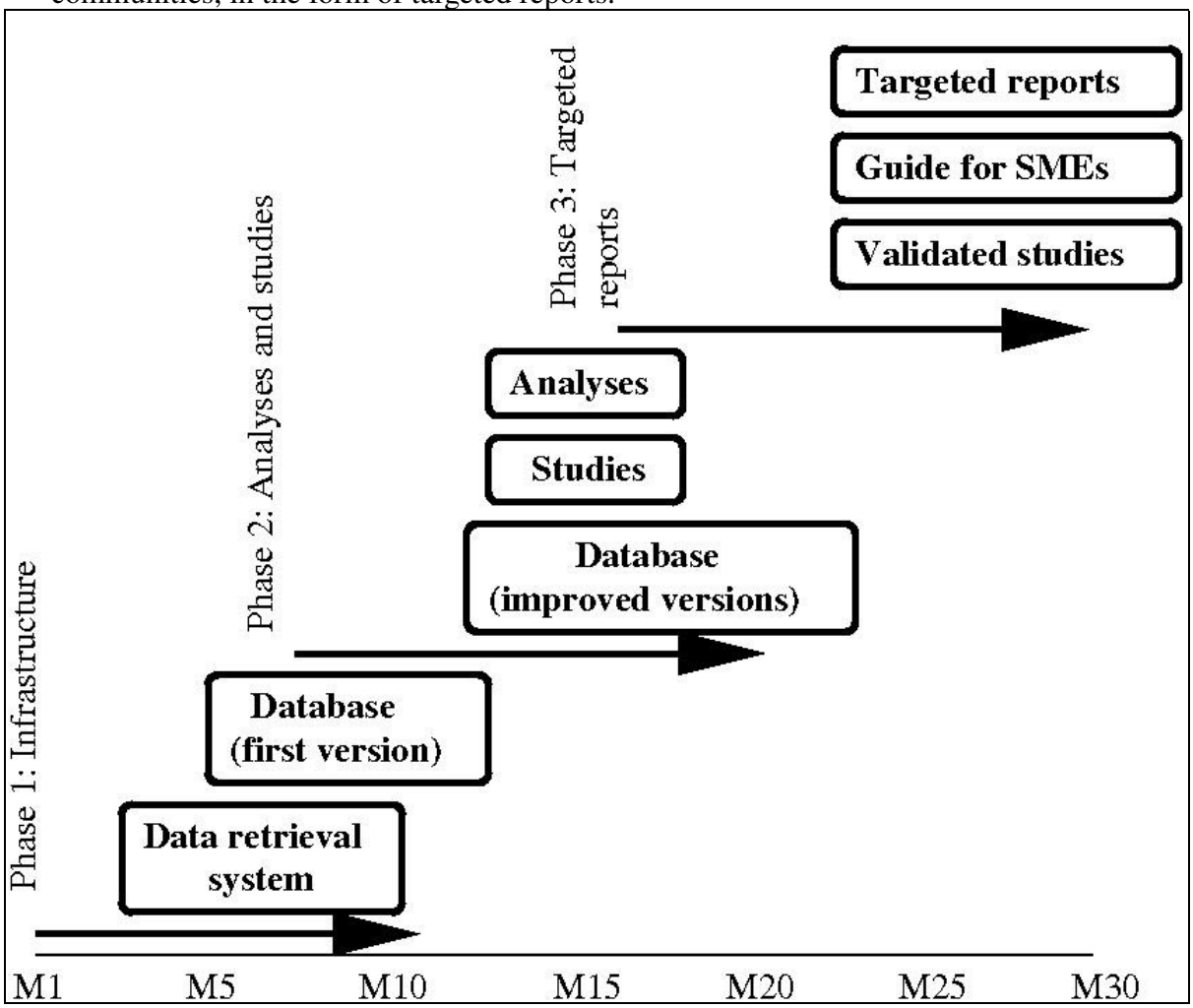
The main outcomes of the project will be:

- Datasets about libre software development, including detailed quantitative information about thousands of projects. These datasets will be stored in a database, available through the Internet.
- A guide on libre software development for SMEs, applying the results of the project to the needs of SMEs interested in libre software.
- Several studies and analysis on different aspects of libre software development, which will enlarge the body of knowledge in this area, and which would be of interest not only to libre software, but to software development in general.
- Several dissemination reports, intended for wide audiences, which can help to better understand scientific facts about libre software development.
- Exploitation guide, which will include recommendations and guidelines for the software industry and for libre software developers

Over time, the main phases of the project would be as follows:

- During a first phase which will cover about the first 10 months, the infrastructure of the project will be designed and deployed, with the focus set into the working of the machinery for feeding the database with information from the repositories of the thousands of selected libre software projects.
- During the next 15 months, most of the analysis and studies will be performed, using also their results and the experience in obtaining them to polish the design of the database and the data retrieval system, so that the data in it is more useful for researchers, and more reliable.

- During the last 15 months of the project (partly in parallel with the previous phase), the results, studies and analyses will be validated and adapted to the needs of the target communities, in the form of targeted reports.



Graphical summary of main outputs and phases of the project