Measurement and Model Building
Discussion and Summary

Sira Vegas and Vic Basili

1 Introduction

This chapter summarizes the discussions that took place during the *Measurement and Model Building* session of the Dagstuhl seminar on Empirical Software Engineering (ESE). The goal of this session was to address questions concerning two topics: data sharing and effective data interpretation.

The chapter has been organized as follows. Section 2 presents the discussions during the data sharing presentations. Section 3 deals with the discussions of effective data interpretation presentations. Section 4 summarizes the topics that were discussed by parallel groups after the presentations. Section 5 sums up the session outcomes.

2 Data Sharing

This section reviews the discussions that took place during Dag Sjøberg's and Richard Selby's presentations concerning data sharing. These two talks are detailed in the *Knowledge Acquisition in Software Engineering Requires Sharing of Data and Artifacts* and *Data Collection, Analysis, and Sharing Strategies for Enabling Software Measurement and Model Building* chapters of this book, respectively.

2.1 Knowledge Acquisition in Software Engineering Requires Sharing of Data and Artifacts (by Dag Sjøberg)

Building bodies of knowledge by accumulating knowledge from empirical studies requires the sustained efforts of several research groups. To achieve this goal, experimental artifacts have to be shared. However, making material accessible to others may require substantial effort by the creator. How should this creator benefit from such an effort, and how should the likelihood of misuse be reduced to a minimum? This talk explores these issues.

This presentation prompted discussion of the following topics: the definition of the term theory, replications by the same vs. different researchers, why a license is needed, who are the owners of the data, and license expiration.

- *Definition of the term theory.* The term theory is not yet well understood by the community, as illustrated by the fact that several people in the audience asked for an example.
- *Replications by the same vs. different researchers.* A study has shown that more confirmatory results are obtained when a study is replicated by the same researchers than when it is replicated by different ones. The speaker pointed out the possibility of a bias. However, there is no agreement with respect to this issue as

V. Basili et al. (Eds.): Empirical Software Engineering Issues, LNCS 4336, pp. 115 – 120, 2007.

some people in the audience remarked that it could be due to human interaction or implicit context variables that differ.

- *Why a license is needed.* There is no agreement in the community on whether a license is really needed. According to the people who are in favor of licensing, the license proposal is motivated by several issues: wanting to know who is using the data artifacts, and any results obtained from its use, the possibility of some researchers misinterpreting or misusing the data without support from the original researchers. However, one suggestion was that we use the Open Source (OS) model about sharing data and artifacts. The risk of misinterpretation or misuse also exists in OS, and this does not stop them from being open.

  Another suggestion was that we should look at what people in physics and medicine do. But a follow up comment was that each field has its own particular problems and issues and we have to tailor any such rules to the ESE field.
- *Who the owners of the data are.* The speaker questioned researchers' ownership of data. However, it is not clear who the real owners of the data are, as different people suggest different options. Should it be the people who provide researchers with the data? The organization these people belong to? The experimental subjects? The government in the case of government-funded projects?
- *Maintaining the data and artifacts.* The issue was raised that data storage, retrieval, evolution costs are real and there is no source of funding that allows data and artifacts to be maintained, quality assured, etc.
- *License expiration.* Somebody in the audience brought up the timing issue. For how long can the people who borrow the data use it?

2.2 Data Collection, Analysis, and Sharing Strategies for Enabling Software Measurement and Model Building (by Richard Selby)

Successful software measurement and model building require effective strategies for data collection, analysis, and sharing. This talk presents a template for data sharing arrangements, along with its application to two environments.

The discussion during this presentation focused on the following points: why data sharing is a problem, its scope, and students as owners of data. Some issues had already been raised in Dag's talk, and they are discussed here in more detail, specifically the issue of data sharing and data ownership.

- *The problem of data sharing.* Some participants did not understand why data sharing is a problem. It was explained that data sharing entails a number of problems. One of them is that it is very costly to maintain data (and the owners have to maintain it). Another one is that some data is private and cannot be shared. This is the case of the data gathered from a company. To illustrate the problem, two examples were given. The first example was that years ago it was possible to use various forms of student programs obtained for experiments, but now these programs are considered the intellectual property of the student and cannot be easily shared. The second example was that the data used to develop various models like COCOMO are often company owned and cannot be shared.

It was also explained that the point is not to obstruct the work of the people who are borrowing the data, but to restrict what they can do with it, and acknowledge the people who own it.

However, some people questioned how real the problem is. They believe that researchers are usually willing to share their data, e.g. the NASA SEL shared its data via a database at the Rome Air Development Center DACS.

- *What the solution is.* Some people doubted that restricting what can be done with the data or acknowledging the people who own it would motivate data owners to lend it. Other proposals were made, like co-authorship of papers (to reflect joint work).
- *Scope of the problem.* Another issue that came up was whether sharing issues affect just data or also extend to any kind of artifact used to run empirical studies.
- *Students as owners of data.* Some people believe that when running experiments with students, they are the owners of their data. In that particular case, we should ask them whether they mind the whole ESE community using their data.

3 Effective Data Interpretation

This section summarizes the discussions that took place during Jürgen Münch's and Audris Mockus' presentations about data interpretation. These two talks are detailed in the *Effective Data Interpretation* and *Software Support Tools and Experimental Work* chapters of this book, respectively.

3.1 Effective Data Interpretation (by Jürgen Münch)

Drawing useful conclusions from individual empirical studies and combining results from multiple studies requires sound and effective data interpretation mechanisms. This talk sketches the progress made in data interpretation in the last few years and presents needs and challenges for advanced data interpretation.

Only one discussion topic came up in this presentation. It focused on:

- *Integrating project monitoring into empirical studies.* The results obtained from project monitoring activities cannot serve as a substitute for running empirical studies. However, project monitoring can be used to help interpret the results of empirical studies. The data from the empirical studies can be used to calibrate models and run simulations.

3.2 Software Support Tools and Experimental Work (by Audris Mockus)

Using software support tool repositories to explain and predict phenomena in software projects is a promising idea. This presentation outlines the opportunities and challenges of using project support systems in empirical work.

The discussion in this presentation focused on understanding data in OS repositories and the role of OS in ESE:

- *Understanding data in repositories.* Although OS repositories are a potential source of data, somebody asked whether it is possible to understand this data. The

answer was that e-mails are particularly useful in these projects. OS people do not sit and talk, but they do everything via e-mail. Therefore, a lot of information can be gathered from e-mails. However, you still need to talk to them, although less so than in closed environments.

- *The role of OS in Empirical Software Engineering.* Somebody asked if it was being suggested that future empirical studies should be derived from OS projects. The answer was no. It is just that OS repositories are a rich source of data, often containing more information than closed environments.

4 Discussion and Summary

The topics proposed for further investigation in parallel working groups were:

- Potential of OS systems as project repositories for empirical studies.
- Documenting theories.
- Data sharing enabling technologies.
- Licensing and sharing issues of qualitative data.
- Prescribing some sort of format for project data sets.

After the voting, the fourth topic was merged with the third one, and the fifth was omitted. This led to the formation of three parallel working groups.

4.1 Potential of OS Systems as Project Repositories for Empirical Studies

The results from the working group examining OS systems as potential project repositories were presented by Audris Mockus and are summarized in the *Potential of OS Systems as Project Repositories for Empirical Studies* chapter in this book.

Only one discussion topic was raised during this presentation, related to the role of the empiricist in the OS project.

- *The role of the empiricist in the OS project.* It was originally suggested that an empiricist should join an OS project as a regular team member. However, somebody asked whether (s)he should not join simply in the role of an empiricist instead. It is possible to join the team as an empiricist as long as you do not bother people with too many questions. But the reason for joining as a team member is that there are always outstanding tasks that an empiricist can easily do (documentation issues, etc.) to help the team out. This way the empiricist is not viewed as an intruder.

4.2 Data Sharing Enabling Technologies

The results of the working group looking at sharing issues were presented by Marvin Zelkowitz and are summarized in the *Ways to share data* chapter in this book.

There was no discussion in the strict sense during this presentation. However, two remarks were made:

- VTT will make all project-related information (including videos, etc.) available via web (for six months)

- It was suggested that this topic should be added to the next ISERN agenda, and the suggestion was accepted.

4.3 Documenting Theories

The results from the working group investigating theory-related questions were presented by Dag Sjøberg and are summarized in the *Documenting theories* chapter in this book.

The topics of discussion raised during this presentation were related to: documenting a theory and the scope a theory should have:

- *Documenting a theory.* Currently there are a lot of theories, but many are still missing. We should know not only what our own theories are, but also how to document them. The http://www.istheory.yorku.ca/ web page was presented as an example of the contents of a theory. The question is whether we should use a similar format. There was agreement on the usefulness of the table associated with each theory, but some people disagreed on whether we could represent our theories using the same format. There was also agreement that we need a way to represent what we currently know, and the table could be used as a guide.
- *Scope of theories.* Another problem is how to articulate theories with respect to their scope. Some people agree that current theories are too specific. Unless you narrow the task, they are useless. These specific theories are not useful for generic phenomena. On the other hand, theories that are generic turn out to be vague. The point is that we should be able to break down theories to the level of abstraction necessary to avoid misfit. Perhaps we are mixing the generic with the specific in our theories. Therefore, we need to examine our theories one by one and decide whether or not they are useful. Somebody pointed out that vague theories are also useful, as they can help to unify terminology.

5 Conclusions

By way of a conclusion of this session, we are going to present the main achievements, key points of dissent and key issues to be solved in the coming years. They are summarized in Table 1.

The **main achievements** identified are related to the findings about the need to share data and artifacts to build bodies of knowledge and theories, OS as an opportunity to study project data under certain circumstances and theories as a way of integrating and motivating empirical studies.

The **key points of dissent** identified during the session are related to the data ownership topic. Part of the community does not believe it is a real problem, and therefore they do not see the usefulness of licensing. Finally, there is no agreement on who the owners of the data are.

Finally, several **key issues to be solved in the coming years** have been identified. They are related to the definition, scope and documentation of theories, how to get

Table 1. Conclusions from the session

	TOPIC	EXPLANATION
Main achievements	Share data and artifacts	We need to share data and artifacts to build bodies of knowledge and theories
	Use Open Source	OS is an opportunity to study project data when there is a good email trail
	Build theories	Theories offer a way of integrating and motivating empirical studies
Key points of dissent	Data sharing	Is it a real problem?
	Data ownership	Who are the owners of the data?
	Licensing	Are licenses really needed?
Key issues to be solved in the coming years	Theories	Definition, scope and documentation
	Replication	How to get confirmatory results in replications by different researchers
	Data ownership	How to acknowledge ownership
	Understanding data	Describe data in a repository
	Maintaining data/artifacts	How do we find the funding to maintain data and artifacts for sharing?

successful replications, how to acknowledge owners of data in a paper, how to describe the data contained in a repository so that it can be easily understood by everyone who uses it and, finally, how to find the funding to maintain data and artifacts for sharing.