Data Sharing Enabling Technologies
Working Group Results

Marvin V. Zelkowitz

Attendees: M. Zelkowitz (chair), V. Basili, R. Glass, M. Host, M. Mueller, T. Ostrand, A. Rainer, C. Seaman and H. Sharp

Issues

If empirical software engineering is to prosper as a research domain, the field needs a mechanism for the sharing of data and artifacts developed by one group that may be useful to another group desiring to work in the same area. In the opening session, the workshop discussed a forthcoming position paper entitled "Protocols in the use of Empirical Software Engineering Artifacts" [1]. This paper describes a taxonomy of properties that are necessary in order to appropriately share information. These properties include who has *permission* to use this data, what *protection* (e.g., privacy) is given to the subjects who provided the data, what *credit* does the user of this data owe the provider of the data, what are the roles governing joint *collaboration* of the activity, who must *maintain* the integrity and access to the data, and what *feedback* does the user of the data owe the creator of the data? The breakout session discussed this taxonomy is greater detail focusing on both qualitative and quantitative data and on what enabling technologies were needed in order to further advance the needs for data ownership and sharing.

The data that is shared can be broken down into 4 classes of artifacts:

1. *Quantitative data* that is the result of measuring an activity. In the software development domain this usually means time data (effort in hours or calendar dates), error data (number and types of defects), and product data (names and sizes of components, execution times, results of testing, etc.)
2. *Artifacts produced* such as source code, design documents, test plans, etc.
3. *Tools* needed to collect data, such as test harnesses, data collection tools, such as Hackystat, etc.
4. *Procedures* for collecting data, such as reporting forms, requirements documents, provided test data, etc.

Qualitative data generally consists of the latter 3 classes. The collected data is often the artifact that is produced. In many cases, such as with ethnographic studies, the collected data consists of interview notes, or tape and video recordings of the subjects of the study.

For both classes of data (i.e., quantitative and qualitative) the proposed taxonomy [1] is still incomplete and still needs additional work in the areas of:

1. *Addressing the proper context of the data.* Understanding the work environment and the proposed application domain for the product under study plays an important role in understanding what the data means.

V. Basili et al. (Eds.): Empirical Software Engineering Issues, LNCS 4336, pp. 108 – 110, 2007.

2. *Understanding common terminology*. Before any meaningful discussion on data sharing can occur, there needs to be a shared understanding of what the objects under study are. During this workshop, three speakers all referenced taxonomies for classifying research methods, and all were different [2, 3, 4].
3. *Need agreement with subjects of what is owned by each*. For example, if data represents a developed program, who owns the rights to that program? Do students own the results of their programming, does the university where the student is enrolled own those programs, or does the research group using the data own those programs? This still is an unresolved issue.
4. *Adherence to national standards and laws*. Related to the previous ownership issue is the fact that rules differ by locality. Various national laws exist governing the collection and dissemination of collected data. It is not obvious how to generate a useful taxonomy usable by the worldwide empirical software engineering community that meets all national and international regulations.
5. *Ethics*. With ownership comes responsibility. What are the obligations and responsibilities of the owner of the data to ensure that the data is used correctly and what are the obligations and responsibilities of the user of the data that results are provided using the proper context for the data?
6. *Provenance*. While maintenance and integrity of the data has already been identified in the taxonomy, the issue of provenance has not been separately identified. As a data set evolves over time, and it will if it represents useful data, then the set of artifacts must be traceable back to their origins and all data must also be accounted for in an unbroken string from its creation to its eventual use.

Enabling Technologies

The focus of the breakout session was to define a roadmap of what enabling technologies, and related research, were needed in order for the data sharing concept to become accepted by the empirical software engineering community. The underlying principle for acceptance of this concept was that "use breeds additional use." If a successful taxonomy can be developed and used by much of the empirical software engineering community, then the rest will follow along. So the issue was how to get an acceptable policy acceptable to most in the field?

The first step is to give the proposed taxonomy wide distribution, with requests for comments and feedback. Copies of the paper [1] were distributed to all workshop attendees and the paper was submitted for publication in *Empirical Software Engineering*. The paper is also scheduled for discussion at the September 2006 International Software Engineering Research Network (ISERN[1]) meeting in Rio de Janeiro, Brazil.

Two other models of sharing were also discussed.

1. *SourceForge.net.*
   a. SourceForge.net, a widely used repository of open source software, maintains a licensing agreement for individuals wanting to use or modify SourceForge products. The success of that licensing policy needs to be studied as an indicator of the problems and issues our empirical software engineering data sharing policy will face.

[1] http://www.cos.ufrj.br/~ght/isern2006.htm

   b. The open source community, as represented by SourceForge, seems to have developed a viable economic model that includes free access to the software. For example, the basic system is free, but add-on features cost extra. This is a major issue in the empirical software engineering data sharing process. We would like data to remain viable for many years, yet it takes funding to maintain the databases. Universities do not have that source of funds and funding agencies are not willing to fund these activities. Perhaps the Open Source model provides a solution.

2. A Material Transfer Agreement (MTA) is a contract that governs the transfer of tangible research materials between two organizations. The MTA defines the rights of the provider and the recipient with respect to the materials and any derivatives. While biological materials are the most commonly transferred items, MTAs may also be used for other types of materials, such as some types of software.

A proposal was discussed covering the ethics issues mentioned earlier. Any paper submitted to a journal using data from an existing source would have an additional review by the creator of that data to ensure that the data was used correctly. This additional review would not be an accept/reject decision (after all, the new paper may correctly say something negative about the original data source, which the additional reviewer may not appreciate), but would provide the editor with additional non-binding comments about the appropriateness of the data analysis used in the paper.

In addition, it was discussed that conferences and journals should provide a small amount of space (and time at a conference) to describe a data source so that others interested in using that data has the opportunity to learn about it and secure a copy. The paper should describe the data set, what it contains, the various constraints in the data, and various experiences that others have had in using the data. The more these are described at meetings, the bigger the market will grow in using data.

We have anecdotal information that advertising data sets do get them used. In revising the 1998 survey on validating computer technology [4] to include data from the years 2000 and 2005 for this workshop, it was noticed that the number of papers using existing data sets greatly increased over the 1998 survey. Most of this increase was in papers using Open Source libraries for such products as Eclipse, the Apache web serve and the Mozilla browser. Given a reliable source of data, researchers are very willing to obtain it for their own research. Formalizing the process with an evaluated taxonomy can only help the process of increasing the supply of good experimental data.

References

1. V. Basili, M. Zelkowitz, D. Sjøberg, P. Johnson and T. Cowling, Protocols in the use of Empirical Software Engineering Artifacts (submitted for publication).
2. D. I. K. Sjøberg, J. E. Hannay, O. Hansen, V. B. Kampenes, A. Karahasanovic, N-K Liborg and A. C. Rekdal, A survey of controlled experiments in software engineering, *IEEE Trans. on Soft. Eng.* 31, 9 (2005) 733-753.
3. W. F. Tichy, P. Lukowicz, L. Prechelt, and E. A. Heinz, Experimental evaluation in computer science: A quantitative study, *J. of Systems and Software* 28, 1 (1995) 9-18.
4. M. V. Zelkowitz. and D. Wallace, Experimental models for validating computer technology, *IEEE Computer* 31, 5 (May, 1998) 23-31.