Data Quality Challenges in Empirical Software Engineering: An Evidence-Based Solution

Michael Franklin Bosu
SERL, School of Computing and Mathematical Sciences
AUT University
Auckland, New Zealand
michael.bosu@aut.ac.nz

***Abstract*—Empirical software engineering data sets are characterized by data quality problems such as noise, outliers, missing data and redundancy. In this paper I propose to address these and other data quality challenges by developing and employing a provenance software tool that is able to explain and replay data capture and processing activities, and to inform the development of appropriate preventive solutions.**

Keywords—data quality; provenance; model; empirical software engineering

I. INTRODUCTION

The management of software development and maintenance projects is guided by models that predict factors such as cost, schedule and quality. Managers therefore rely heavily on the accuracy of these models, and in turn model accuracy is reliant on the data and methods used in their development. While extensive research has been conducted in the area of model development, the same cannot be said regarding the quality of data that serve as input to these models. Data quality challenges such as noise, outliers, missing data, redundancy, small data set size, and data accessibility have been identified in empirical software engineering (ESE) data sets [1–3].

The aim of this research is to augment software development management environments so that software systems are made provenance-aware. As a result, ESE data will be accompanied by provenance data, which will offer researchers and practitioners the opportunity to assess the quality of data and also increase the trust associated with data sets. This will improve the practice of software engineering as data collection procedures are typically not reported in the ESE domain. On the few occasions that data collection has been reported, problems associated with data collection were not discussed.

The rest of this paper is organized as follows. In section II, I present related work with an emphasis on the data quality challenges evident in ESE. In section III, I present the research approach I intend to use to achieve the research objectives, and in section IV, I present the progress made to date.

II. RELATED WORK

Measurement data are used to support many aspects of software development and management, but effort estimation and defect classification are particularly prevalent uses of such data. The notion of quality is often defined as signaling fitness for purpose [4] and the expectation is for ESE data to be of high quality – to be fit for the modelling task at hand, be it for classification or prediction. The empirical software engineering community is unfortunately not immune to studies that have used questionable or poor quality measurement data in model-building [1–3]. As a result, the quality of empirical software engineering data sets has been subjected to increasing scrutiny. For instance, Gray et al. [5] discovered several data quality problems with the NASA Metrics Program data sets that are widely used for defect prediction. These problems, which included redundant data, inconsistencies, constant attribute values, missing values and noise, are commonly noted in other empirical software engineering data sets. According to Liebchen and Shepperd [6], data quality has not received sufficient attention from the software engineering community as evidenced by the few studies that explicitly report the quality of data sets.

The research community has responded with several techniques to combat some of these challenges. Imputation techniques, for example, have been introduced to address the issue of incomplete data [7]. Noise identification and filtering techniques were applied by several authors [1], [8] to empirical software engineering data sets to improve data quality. A detailed study conducted by Bosu and MacDonell [9] classified all the major data quality challenges in ESE into a taxonomy and identified the class of provenance as the one that had received the least attention in the ESE domain (when compared to issues of accuracy and relevance). Provenance in the taxonomy relates to accessibility and trust issues concerning ESE data sets. Measures that have been used in addressing data quality challenges were also identified in the taxonomy.

To the best of our knowledge, there is only one work on provenance in the ESE domain[10] which considered the provenance of software entities, specifically, the introduction of an anchored signature method to determine the provenance of source code contained within Java archives. We believe that the provenance of data about the source code is as important as the provenance of the source code.

The major (long-term) goal of this research is to improve the quality of data sets used in empirical software engineering. Although several mechanisms have been tried to resolve the challenges associated with data quality, currently there are no best practices to follow to address any of these problems. It is argued here that this is due to the lack of a clear understanding of the causes of these problems. The focus of this research is on

providing a provenance solution that is able to explain the causes of the problems of data quality, thus offering researchers and practitioners a clearer understanding of the root causes of problems which in turn will lead to better ways of addressing these challenges. The trust associated with data sets will improve as stakeholders will have access to provenance information to enable the full assessment of data quality - this in turn will lead to the overall improvement of software engineering practice. The following are the research questions to be addressed:

1. What processes and procedures do organisations use in measurement data collection and processing?
2. Why are there trust issues with data sets in empirical software engineering?
3. What parts of a measurement system records provenance, or need to be adapted in order to record provenance?

III. Proposed Research Approach

To address data quality problems the intent is to take a holistic approach by attempting to develop a solution that can be applied at the data capture and processing stages. This will leverage knowledge from provenance systems to provide a provenance solution that has the potential of explaining the causes of quality problems. Specifically, the Provenance incorporating methodology (PrIMe) [11] will be employed to make an existing measurement system provenance-aware. This is a research with a design component and as such the design science research process proposed by Hevner et al. [12], which has been widely used in software engineering and information systems research, will be utilized. We will also use the Nunamaker et al. [13] System Development Research Methodology in designing and constructing a software tool.

This research will address problems of data quality by employing a provenance solution. The notion of data provenance relates to the trustworthy and auditable recording of information regarding the source and processing of data. The intent is to follow a constructivist research approach to design, develop and evaluate a software system that is able to capture provenance information throughout the data collection and processing stages in the context of software engineering measurement. This system will also have the potential of replaying provenance information that could be used to explain problems associated with data should they arise. The system will be evaluated in two ways: first, its core functionality will be evaluated based on laboratory experiments; second, it will be evaluated for utility by relevant experts.

This research will also develop a data quality assessment model and an auditing model that will serve as a means of assuring the trustworthiness of ESE datasets. These models will be evaluated based on provenance data.

The novel contributions of this research will be a provenance framework comprising the provenance system, a trust model, a data quality assessment model and an auditing model for empirical software engineering data.

IV. Research Progress

A comprehensive literature review has been conducted which identified all the major challenges associated with ESE data quality and the state of practice of data quality by the ESE community. Hackystat, a framework that provides facilities for the collection, analysis, visualization, interpretation, annotation, and dissemination of software development process and product data, is the measurement environment that will be made provenance-aware. In order to make the solution more general, the data collected from Hackystat will be that which are common to most measurement systems. We have identified some of the provenance use cases and actors that are essential in obtaining provenance information. Analysis of Hackystat for provenance use cases and actors is ongoing. I am also in the process of designing the architecture of the system and settling on the provenance storage design to use to hold the provenance data. I have identified the necessary tools (Java) and libraries (PreServ, Client Side Library) that will enable me to build the solution. I will then make adaptations to the generic provenance architecture to develop a tailored software tool that is able to capture provenance information and supports queries to answer provenance questions.

References

[1] J. Van Hulse, T. M. Khoshgoftaar, C. Seiffert, and L. Zhao, "Noise Correction Using Bayesian Multiple Imputation," Information Reuse and Integration, 2006 IEEE International Conference on Information Reuse and Integration, pp. 478–483, 2006.

[2] P. M. Johnson and A. M. Disney, "A Critical Analysis of PSP Data Quality : Results from a Case Study," Empirical Software Engineering, vol. 4, no. 1, pp. 317–349, 1999.

[3] G. Liebchen, B. Twala, M. Cartwright, and M. Shepperd, "Assessing the Quality and Cleaning of a Software Project Data set : An Experience Report," 10th International Conference on Evaluation and Assessment in Software Engineering (EASE), pp. 1–7, 2006.

[4] P. B. Crosby, Quality is free: The art of making quality certain. McGrawHill, 1979.

[5] D. Gray, D. Bowes, N. Davey, and B. Christianson, "The misuse of the NASA Metrics Data Program data sets for automated software defect prediction," 15th International Conference on Evaluation & Assessment in Software Engineering (EASE 2011), pp. 96–103, 2011.

[6] G. A. Liebchen and M. Shepperd, "Data sets and data quality in software engineering," Proceedings of the 4th international workshop on Predictor models in software engineering - PROMISE '08, pp. 39–44, 2008.

[7] T. M. Khoshgoftaar and J. Hulse, "Imputation techniques for multivariate missingness in software measurement data," Software Quality Journal, vol. 16, no. 4, pp. 563–600, Jun. 2008.

[8] T. M. Khoshgoftaar and J. Van Hulse, "Identifying Noise in an Attribute of Interest," Proceedings of the Fourth International Conference on Machine Learning and Applications, pp. 55–62, 2005.

[9] M. F. Bosu and S. G. Macdonell, "A Taxonomy of Data Quality Challenges in Empirical Software Engineering," 22nd Australasian Software Engineering Conference, in press.

[10] J. Davies, D. M. German, M. W. Godfrey, and A. Hindle, "Software Bertillonage : Finding the Provenance of an Entity," In Proceeding of the 8th working conference on Mining software repositories., pp. 183–192, 2011.

[11] S. Miles, P. Groth, S. Munroe, and L. U. C. Moreau, "PrIMe : A Methodology for Developing Provenance-Aware Applications," ACM Transactions on Software Engineering and Methodology (TOSEM), vol. 20, no. 3, 2009.

[12] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design Science in Information Systems," MIS Quarterly, vol. 28, no. 1, pp. 75–105, 2004.

[13] J. F. Nunamaker, M. Chen, and T. D. M. Purdin, "Systems Development in Information Systems Research," Journal of Management Information Systems, vol. 7, no. 3, pp. 89–106, 1991.