Achieving Reproducibility Incorporating Service Versioning into Provenance Model

Dayang Hanani Abang Ibrahim[1], Nadianatra Musa[1], Chiew Kang Leng[2], Jane Labadin[2], Johari Abdullah[3], Sarina Sulaiman[4]
[1]*Department of Information Systems,*
[2]*Department of Computational Science and Mathematics,*
[3]*Department of Computer System and Communication Technology,*
Faculty of Computer Science and Information Technology,
Universiti Malaysia Sarawak, Kota Samarahan, 94300 Sarawak, Malaysia.
[4]*UTM Big Data Centre, Universiti Teknologi Malaysia, Skudai, 81310 Johor, Malaysia.*
hananii@unimas.my

***Abstract*—Reproducibility has long been a cornerstone of science. Underpinning reproducibility is provenance, which has the potential to provide scientists with a complete understanding of data generated in e-experiments, including the services that were produced and consumed. This paper explores the issues of service versioning in provenance to achieve reproducibility. Current provenance model does not directly support service versioning. Therefore, this paper introduces an enhancement of a provenance model to incorporate service versioning mechanism that provides a way to access multiple versions of the same service so that researcher can compare one version to another, and understand their effects on processing data. The enhanced provenance model is able to track the changes of the same service (versions of the same service) over time and correlates versioned services with the results they generate.**

***Index Terms*—Reproducibility; Provenance; Provenance Model; Service Versioning; Web Services Architecture;**

I. INTRODUCTION

Provenance is particularly important when a scientific e-experiment is to be reproduced and re-run. Provenance provides the ability to reproduce all the steps leading to a scientific e-experiment result. This means provenance can show how the result was generated, thus illustrating how the experiment was done before. Pizzi et al. [1] uses directed acyclic graphs to track the provenance of data and calculations in computational science to ensure reproducibility. A service is a unit of work that performs a computation that can be consumed by clients or consumers in applications or experiments. When a workflow is executed, a sequence of services is invoked. Provenance enables the recording of these services, including the data parameters used, and also timestamps of service invocations. Looking inside each of these services, there are also service metadata that may also be significant and therefore needs to be recorded in provenance; for example, when a particular service was created and which version it is. In existing provenance literature, versioning has not been directly supported in provenance model. It is often the case that a service will need to change after its initial deployment to fix bugs, improve the algorithm, or meet new requirements. Therefore, service versioning should be supported to ensure that even after new versions of a service are deployed; the old version still remains available. This evolution of services will eventually lead to multiple versions of a service, starting with the current version, and leading back to older versions that have in the past been used to generate data that may still be in use. This piece of service metadata is important for reproducibility. Therefore, reproducibility not only gives relevant information to permit the re-running of the experiment but also to look at the versions of a service that have been invoked in the experiment. This approach opens up the opportunity for discovery in examining the history of the service. As researchers have realised that reproducibility can promote sharing, and give other advantages to the scientific community, there has been a growth in work on reproducibility [2][3][4]. These works discuss the motivation for reproducibility, as well as describing infrastructure to support it.

Experimental reproducibility is concerned with being able to re-execute past experiments in a different workflow environment and to see if a prior result can be confirmed. This is because it is not guaranteed that past experiments can be re-executed successfully if the experiments were created in a different workflow environment. This may due to a different workflow structural differences and missing data, services or processes. To reproduce experiments, the original experimental entities must be accessible. To achieve this, reproducibility requires provenance information that captures all the important entities in an experiment. For this to be successful, the entities must be described by a provenance model. A major issue is that the experimental entities may be changed from time to time: for example, new versions of services used in an experiment may be deployed. Therefore, in this paper we argue that versioning is an essential mechanism needed to support experimental reproducibility.

Over the years, the research community has realised that a major problem in sharing its research experiments with others, is the inability to reproduce past experiments. This problem is caused by i) insufficient information describing the experiment and ii) research (experimental) artifacts and processes (services) that are not available.

This reproducibility process therefore needs provenance information to describe the execution of the experiment in a way that can allow reproduction. In addition, the experimental artifacts and services should be made accessible for later use. Therefore, the essential concepts underlying the reproducibility of experimental results are capturing the computation, along with the data on which it operates. In service-based e-science, the fundamentals of a computation