

Squinting at the data:
Investigating software artifact provenance using KISS techniques

Michael W. Godfrey
University of Waterloo
Visiting Distinguished Scientist at CWI-Amsterdam

``Provenance'' is a term from archaeology and the arts that refers to a set of evidence supporting the claimed origin of an artifact, such as a piece of pottery or an oil painting. Recently, the term has been used in an electronic context --- ``digital provenance'' --- to indicate an artifact such as a software component or set of data, really is what it claims to be and should be permitted to be used within sensitive operating environments.

In this talk, I suggest how we can stretch the definition further to encompass "software artifact provenance". That is, for a give software development artifact such as a user-visible feature, a source code function, or a third-party library, we might want to ask the question: Where did this come from and what is the evidence? For example, one might wonder how a given feature was decided upon during a mailing list discussion, how it manifested itself in the code, and how it has been maintained since the initial implementation. For a given function, one might wonder about its history within the design of the system: Was it designed to fit exactly here, or was it moved or cloned from elsewhere? And for a given third-party jar file that has been included in a Java system distribution, one might ask: What version of the library is this, and how do we know?

In this talk I will sketch some of the ideas behind this work, and show how we might phrase some of these questions in terms of concrete criteria. In particular, we will concentrate on simple techniques for reducing a large search space of candidates down to a small handful that can be examined in detail using more expensive techniques. A concrete example of investigating third-party libraries in Java systems will be presented.

This is joint work with Daniel German of the University of Victoria, Julius Davies of the University of British Columbia, and Abram Hindle of the University of Alberta.