Bridging the Gap between Process and Procedural Provenance for Statistical Data

George Alter*[1], Timothy Mcphillips[2], Thomas Thelen[3], Jack Gager[4], Bertram Ludäscher[2], Dan Smith[5], and Jeremy Iverson[6]

[1]University of Michigan – United States
[2]University of Illinois at Champaign-Urbana – United States
[3]University of California at Santa Barbara – United States
[4]Metadata Technologies North America – United States
[5]Colectica – United States
[6]Colectica – United States

Abstract

We show how two models of provenance can work together to answer basic questions about data provenance, such as "What computed variables were affected by values of variable X?" The W3C PROV data model is a standard for describing activities and persons that produce digital artifacts. PROV associates processes with inputs and outputs, but it does not have a way to describe how data are changed within the process. PROV has no language for program components, like mathematical expressions or joining data tables. Structured Data Transformation Language (SDTL) provides machine-actionable representations of data transformation commands in the five most widely-used statistical analysis applications. SDTL is a procedural language in which commands are executed sequentially. Thus, SDTL describes the inner workings of programs that are black boxes in PROV. However, SDTL is detailed and verbose, and simple queries can be very complicated in SDTL. Combining PROV and SDTL allows us to answer questions about data preparation and management at levels not available in PROV. Our bridge between PROV and SDTL rests on two pillars: ProvONE, an extension of PROV, and Structured Data Transformation History (SDTH), a simplified view of SDTL.

*Speaker