Quantitatively Understanding Workflow Performance using `prov:Bundle`

Eric Stephan, Todd Elsethagen, Kerstin Kleese van Dam, Bibi Raju, Alok Singh, Ilkay Altintas, Darren Kerbyson

Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

The notebook is the place where all primary data must be recorded. Paper towels, napkins, toilet tissue, or scratch paper have a tendency to become lost or destroyed. It is a bad practice to record primary data on such random and perishable pieces of paper. [1]

CONCEPT

- Yes even 40 years later[1] non-traditional forms of provenance are still very prevalent!
- Using non-traditional forms may be motivated by
  - legacy practices
  - Innovation
    - Allowing new ways to analyze and explain data origin, process history
    - Providing efficiency in the way provenance can be stored and managed.

NEW IDEAS

•`Prov:Bundle` recognizes the co-existence of non-traditional forms of provenance with PROV:

- Supports provenance of provenance: Provides the means to associate different kinds of provenance as an object to PROV graphs.

•Definition offers some flexibility, *leaving it open to different interpretations*…Provenance of Provenance as a(n):

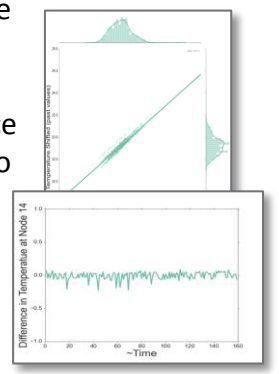
- Native source provenance `dcat:Dataset`
- Subset persisted in an external store.
- Algorithm

IMPACT

- Including additional crucial historical facts even in non-traditional forms, help explain results.
- Technical barriers can be reduced in extreme scale applications by introducing innovative approaches to provenance representation.
- Raw chain of custody evidence or domain oriented provenance in addition to PROV benefits consumers.
- Audit trails associated with provenance can help identify personal data For example, fingerprinting an integral part of provenance and needs to be managed.

MOTIVATION

- Native source: Earth system modelers compile case study provenance information into their application as a dataset.
- Hybridization: Studying workflow performance by streaming time-series metrics correlated to workflow provenance to effects of workflow task
- Using empirical data collected from provenance and metrics to support adhoc decisions made by trained algorithms.



[1] Pavia, Donald L. Lampman, Gary M. Kriz, George S. Donald L. Pavia, Gary M. Lampman, and George S. Kriz. Introduction to organic laboratory techniques: a contemporary approach. 1976.