CPR
Transparency for Whole Tale Recorded Runs

Timothy M. McPhillips, Thomas Thelen, Craig Willis,
Kacper Kowalik, Matthew B. Jones, and Bertram Ludaescher

21-July-2021
Provenance Week 2021

NSF Award OAC-1541450

Publishers and conference organizers **increasingly verify** the reproducibility of **computational artifacts**.

- Whole Tale (WT) is ideally suited to support verification workflows.
- A *Tale* is a computing environment shared between authors and verifiers.
- WT provides seamless integration with data repositories.
- WT supports MATLAB, STATA, and other runtimes that require licensing.

Computational **reproducibility platforms today** best **support black-box verification**, i.e. **confirming a** computational **recipe** as a whole **produces** the **expected results.**

- A Tale declares a recipe for a computing environment, references to datasets used by the Tale, and an entry point for the computation.
- A Tale currently does not report what components are used to produce which artifacts or how.

Successfully **repeating** a **computation** based on a description of it **is a test of transparency**; it **shows** that the **description** of the computation **was** sufficiently **complete and correct**.

- Verification workflows currently answer two primary questions:
- Is the description provided in the paper sufficient to enable others to repeat the reported computations?
- Does repeating the computations yield the reported results?

The **R-word** most **fundamental to** the practice of modern **science is** neither ‘reproducibility’ nor ‘replicability’. It is ‘**reviewability**’. In its absence the other R-words are of little value.

A completely **invalid computation can be** perfectly “**reproducible**”.

Whole Tale's **recorded run*** feature **represents** a **transparent-box** approach to **verification** workflows and reproducibility.

- The *Comprehensible Provenance Record (CPR)* toolkit* transforms raw provenance records into meaningful artifacts.
- CPR employs ReproZip to monitor system calls.
- Detects which programs and libraries installed in the Tale container are actually used.
- Detects which programs read or write each data file.

**Currently under development.*

Recording how **data flows** between programs invoked during a Recorded Run will **reveal** the **distinct provenance of different** computed **products**.

- Enables verifiers to pose a new question:
- Is a paper's description of the roles played by software components consistent with the observed flow of data?

Predefined **queries**, **reports**, and **visualizations** **make provenance useful** to verifiers and others trying to understand a paper and its results.

- CPR converts ReproZip traces to RDF triples.
- The CPR vocabulary extends PROV and ProvONE.
- Standard reports and visualizations declared using *Geist* highlight data flow.
- Parameterized queries yield lineage of particular products.
- Eager reasoner in Blazegraph enables each SPARQL query to employ the most convenient terms from multiple vocabularies.

Observing system calls alone cannot possibly yield all of the provenance information necessary to **make** computations or **research transparent**.

- CPR currently cannot see activity inside scripts.
- Cannot see inside compiled programs.
- Cannot see inside Jupyter notebooks or R Markdown.
- Cannot see inside remote web services.
- Cannot see inside workflow management systems.

To make computational elements of **research** studies truly **transparent**, **provenance must be captured** automatically **at multiple levels by multiple means** that integrate robustly.

- Whole Tale aims to incorporate additional provenance recording methods.
- Provenance queries will integrate provenance from different sources.
- **Provenance features developed for Whole Tale also will work outside of Whole Tale**.

Please come to the T7 Workshop on Provenance for Transparent Research!

- Hear short talks like (but much better than) this one.
- Vote on whether assertions made by speakers align with your own experience and research disciplines.
- Share your own insights and get feedback from others.
- Connect with domain researchers who want to incorporate provenance management into their own tools and fields.

We hope to see you at 9 AM UTC-5 Thursday