Adventures in Provenance Integration

Elaine Angelino, Uri Braun, David A. Holland, Peter Macko, Daniel Margo, Margo Seltzer

TaPP 2011
Integrating two provenance systems

• StarFlow, a workflow environment for data analysis in Python

• PASS, which manages provenance at the operating system level

• Run StarFlow on top of PASS
The dream

We thought that PASS and StarFlow would get along like two peas in a pod ...
The reality

... But sometimes they were more like apples and oranges
Reconciliation

• StarFlow and PASS each has its own world view
  – Different scopes, objects, granularity, goals
• Each produces its own account of any particular execution
• These accounts need to be reconciled semantically as well as physically
Non-existent objects

- Problem: StarFlow records information about objects that do not yet exist
- The OPM cannot express this

- Solution: Placeholder objects
- The new *stands-for* edge connects placeholders to reality
Version disconnection

- Problem: When StarFlow regenerates a file, it deletes the old file and writes a new copy
- PASS sees no connection between the old and new objects

- Solution: Explicit version edges
- StarFlow must tell PASS about these relationships
Problem: We let StarFlow add provenance records to the PASS database

PASS needs to keep track of where individual records came from

Solution: ???

What about the provenance of provenance of provenance of provenance of provenance of provenance …
Is this what provenance integration looks like?