Towards Verifiable Publications
Tanu Malik
Computation Institute
University of Chicago and Argonne National Laboratory
Scientific Publication Cycle

- Data Collection, Research & Analysis
- Authoring
- Review
- Publication & Citation & Impact

Participants:
- Researchers
- Authors
- Reviewers
- Readers
Scientific Publication Handoffs

- Researchers
  - Data, Computation, Software, Hardware

- Authors
  - Latex, Word

- Readers
  - PDFs

- Reviewers
  - PDFs
Impedance Mismatch, Lack of Verification

Computational science inputs are not linked with outputs.

- **Inputs**: Large quantities of data, complex data manipulation and/or numerical simulation use of large and often distributed software stacks, etc. (software, data, execution, environment)

- **Outputs**: Research papers (text-based, non-interactive)
Verifiable Publications

- Verifiable publications: Augmenting publications of with the computer code used to carry out the research, such that readers/reviewers can
  - repeat (or replicate) results on nominally equal configurations,
  - generalize the results by replaying them on new data sets, verify how they vary with different parameters,
  - re-use and extend the experiment.
Science Object Linking and Embedding

• a framework for creating and supporting verifiable publications

• consists of a set of easy-to-use tools for creating science objects

• science objects are self-contained software or data artifacts that can be linked to a publication

• help readers/reviewers to repeat, reuse, or repurpose
SOLE Science Objects

• Granular units of knowledge that are self-contained
  – Consist of content of a data, method or process, and metadata, such as environment or Uniform Resource Identifier (URI) related to the content.
  – Examples: language objects, annotation objects, software packages, virtual machine images
  – Several tools for encapsulating data, method or process, but lack adequate metadata for verification
SO Creation: Easy 3 step process

Run
- Run SOLE
- SO tool

Add
- Provenance metadata
- Annotation metadata

Share
- Link to Publication
- Re-execute
SOLE Tools

- **Modified Ctags:** User-defined tags on source files to create language objects on the fly
- **PDFExtractor:** Extracts segments, such as text, images that are selected by the user through the annotation command in the PDF reader
- **PTU Package:** Create a provenance-aware, portable software package that requires no installation, configuration, or privilege permissions
- **Light-weight Database Virtualization:** Portable packages for database applications
Language Objects

User-defined tags on source files to create language objects on the fly

Step 1: Tag source file test.R

Step 2: Run SOLE on test.R; Create SO

Step 3: Search SO with tag name and link
Annotated Objects

Add Tags
- Annotate the PDF by selecting a rectangular region

Run PDFExtractor
- Upload tags to SOLE

Share
- Link tags to phrases in publication
other compliance markets (Texas, Maryland, Pennsylvania, and Washington D.C.) due to a continued surplus of eligible renewable energy supply relative to RPS-driven demand. Prices for RECs offered in voluntary markets in 2010 ranged from an annual average of less than $1/MWh for national voluntary wind RECs (which continue to closely track the price of Texas RECs) to approximately $6/MWh for voluntary wind RECs in the West.

The extracted text from the paper (top). The highlighted value $1 per MWh refers to the annotation science object with the tag Texas (bottom).
PTU-Package

- **Provenance-To-Use-Package**
  - Create a portable software package without installation, configuration, or privilege permissions

- **Authors build a package that includes:**
  - Software program
  - Input data
  - Provenance trace

- **Testers may select a subset of the package’s processes for a partial or full deterministic replay**
  - Guarantees that events are processed in the same order using the same data
PTU

Audit Program
- Copy data and source files

Extract Provenance
- Include provenance

Share
- Repeat the package
PTU Functionalities

• **ptu-audit** tool
  – Build a package of authors’ source code, data, and environment variables
  – Record process- and file-level details about a reference execution

• PTU package
  – Display the provenance graph and accompanying run-time details

• **ptu-exec** tool
  – Re-execute specified part of the provenance graph
ptu-audit

- Uses *ptrace* to monitor system calls
  - execve, sys_fork
  - read, write, sys_io
  - bind, connect, socket
- Collects provenance
- Collects runtime information
- Makes package
ptu-package

Provenance DB

/data/conf.dat
/read()

/data/in.dat
/read()

/bin/filter
/write()

/bin/convert

/bin/workflow

/bin/calculate

/bin/reclass

/bin/python

$PKG_ROOT/bin/
workflow
reclass
filter
calculate
convert
python
data/
conf.dat
in.dat
out.dat
out.dat

PTU Package
ptu-exec

TESTER's COMPUTER

PTU

Verify, Start calculate and Redirect

ptu-exec  ptu-audit

Read Graph

PTU Package

Provenance DB

$PKG_ROOT/bin/
calculate
filter
data/in.dat
out.dat

$PKG_ROOT/bin/
calculate
read($PKG_ROOT/data/in.dat)
write($PKG_ROOT/data/out.dat)

$PKG_ROOT/bin/
filter
CDE read()

/data/conf.dat
/data/in.dat
/data/out.dat
Current PTU components

- Uses `ptrace` to create a provenance graph representing a reference run-time execution
- Uses SQLite to store the provenance graph
- Uses `graphviz` for graph presentation
PTU Features

- Spans local and distributed computations
  - Creates different packages for heterogeneous environments
- Requires no installation, works under user context
  - Suitable for authorized, controlled environments
- Limited to Linux environments
  - Works for all distributions of Linux, irrespective of dependency overlap
Working Papers

- [http://www.ci.uchicago.edu/SOLE](http://www.ci.uchicago.edu/SOLE)
- Fifteen examples of research papers from the Center for Robust Decision making on Climate and Energy Policy (RDCEP)
  1. *Workflow papers*: The author must associate the text and embedded figures with science objects that include datasets, algorithmic descriptions, computational analysis workflows, and workflow executions (3 papers)
  2. *Policy paper*: The author must associate descriptions in the paper with a set of data values, each of which is embedded in another research paper. (1 paper)
  3. *Dataset-based papers*: Many papers from PubChem (10 papers)
  4. *Simulation and modeling papers*: Differential equation with parameters obtained from the Web. (1 paper)

Get Data
- Wget
- Bash script

Reclassify
- R
- Raster
- Rgdal
- Rgdal Tools

Calculate
- R
- Geo algorithm
Results

- 3-step workflow with 7 classifications
- 3-step workflow replayed in PTU with 1 classification
TextAnalyzer

- runs a named-entity recognition analysis program using several data dictionaries
- splits the input file into multiple input files on which it runs a parallel analysis
Results

- No Monitoring
- PTU audited for provenance & package
- Replay with lower convergence criteria

<table>
<thead>
<tr>
<th></th>
<th>Real time</th>
<th>User time</th>
<th>System time</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Monitoring</td>
<td>72.035</td>
<td>1.773</td>
<td>2.844</td>
</tr>
<tr>
<td>PTU audited</td>
<td>87.41</td>
<td>32.505</td>
<td>13.408</td>
</tr>
<tr>
<td>Replay</td>
<td>72.035</td>
<td>1.773</td>
<td>2.844</td>
</tr>
</tbody>
</table>
Current Work

- **PTU packages**
  - good for repeatability not reproducibility
  - Adding binary applications is cumbersome
- **Explore virtual machines as another option**
  - Raises some double virtualization issues
- **Improve readability of VMIs**
  - File tags related to paper tags
Acknowledgements

• Collaborators: Quan Pham, Ian Foster
• Victoria Stodden
• Scientists at the Center for Robust Decision Making on Climate and Energy (RDCEP), the Division of Decision Making, Argonne National Laboratory, Institute of Genomics and System Biology, University of Chicago
• Funding: