VisTrails: A Python-based Scientific Workflow and Provenance System

David Koop
University of Utah and VisTrails, Inc.

VisTrails - Visual Workflow Programming
Structure Adds Value

1. <book isbn="0321559673">
   <title>XML: Visual QuickStart Guide</title>
   <author>Kevin Howard Goldberg</author>
   <edition>2</edition>
   <price>23.09</price>
   <rating>5</rating>
</book>

2. <book isbn="0470114878">
   <title>Beginning XML</title>
   <author>David Hunter</author>
   <author>Jeff Rafter</author>
   <author>Joe Fawcett</author>
   <author>Eric van der Vlist</author>
   <edition>4</edition>
   <price>26.39</price>
   <rating>3.5</rating>
</book>

3. Structure Adds Value

```python
import vtk
data = vtk.vtkStructuredPointsReader()
data.SetFileName(../examples/data/head.120.vtk)
contour = ...
renderer = ...
renWin = ...
iren = ...
style = ...
iren.SetRenderWindow(renWin)
iren.SetInteractorStyle(style)
iren.Initialize()
iren.Start()
```
Workflow Differences

Workflow Query-by-example

[IEEE Visualization 2007]
Workflow Completions

Provenance in Art

- **Rembrandt van Rijn**
- **Dutch, 1606 - 1669**
- **Self-Portrait, 1659**
- **oil on canvas**
- **Andrew W. Mellon Collection**
  - 1937.1.72
- **Provenance**
  - [1] This early provenance is established by presence of a mezzotint after the portrait by R. Earlom (1743-1822), dated 1767. See John Charrington, A Catalogue of the Mezzotints After, or Said to Be After, Rembrandt, Cambridge, 1923, no. 49.

- **Associated Names**
  - Buccleuch, Henry, 3rd Duke of
  - Buccleuch, John Charles, 7th Duke of
  - Colnaghi & Co., Ltd, P. & D.
  - Knoedler & Company, M.
  - Mellon, Andrew W.
Provenance in Science

- **Provenance is as (or more) important as the result!**
- Old solution:
  - Lab notebooks
- New problems:
  - Large volumes of data
  - Complex analyses
  - Writing notes doesn’t scale
- New solution:
  - Automated provenance capture with user-defined annotations

<table>
<thead>
<tr>
<th>Date</th>
<th>Annotations</th>
<th>Observed Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>11/20/96</td>
<td></td>
<td>[DNA Recombination, Lederberg]</td>
</tr>
</tbody>
</table>

Provenance in VisTrails

- **Design Provenance (Version Tree)**

- **Execution Provenance (Logging):**

```xml
<module id="12" name="vtkDataSetReader" start_time="2010-02-19 11:01:05" end_time="2010-02-19 11:01:07">  <annotation key="hash" value="c54bea63cb7d91f2a43ce"/></module>
<module id="13" name="vtkContourFilter" start_time="2010-02-19 11:01:07" end_time="2010-02-19 11:01:08"/><module id="15" name="vtkDataSetMapper" start_time="2010-02-19 11:01:09" end_time="2010-02-19 11:01:12"/>
```
Design Provenance

VisTrails - Design Provenance
Specifications

- Open-source, freely downloadable system (www.vistrails.org)
- Multi-platform: users on Mac, Linux, and Windows
- Python code and uses PyQt and Qt for the interface
- Over 13,000 downloads since 2007
- User’s guide, wiki, and mailing list
- Many users in different disciplines and countries:
  - Visualizing environmental simulations (CMOP STC)
  - Simulation for solid, fluid and structural mechanics (Galileo Network, UFRJ Brazil)
  - Quantum physics simulations (ALPS, ETH Switzerland)
  - Climate analysis (CDAT)
  - Habitat modeling (USGS)
  - Open Wildland Fire Modeling (U. Colorado, NCAR)
  - High-energy physics (LEPR, Cornell)
  - Cosmology simulations (LANL)
  - Using tms for improving memory (Psychiatry, U. Utah)
  - eBird (Cornell, NSF DataONE)
  - Astrophysical Systems (Tohline, LSU)
  - NIH NBCR (UCSD)
  - Pervasive Technology Labs (Heiland, Indiana University)
  - Linköping University
  - University of North Carolina, Chapel Hill
  - UTEP

Climate Data Analysis

[CDAT Project, Lawrence Livermore National Lab]
Quantum Lattice Models

Coastal Margin Observation & Prediction
Comparing Cosmological Simulations

Wildfire Prediction

[www.vistrails.org]

[Cosmic Code Comparison Project, Los Alamos National Lab]

[WRF-Fire Project, University of Colorado-Denver]
Extensibility

- Package infrastructure
- Wrap python libraries, command-line calls, or use other interfaces (jpype, rpy, etc.)
- Need to specify:
  1. Package identification information
  2. Module structures: input & output ports
  3. Compute method for each module
Extensibility Example

- seawater python package:
  - [http://pypi.python.org/pypi/seawater/1.0.3](http://pypi.python.org/pypi/seawater/1.0.3)

```python
identifier = 'org.ocefpaf.seawater'
version = '1.0.3'
name = 'Seawater Routines'
import seawater

class SaturationN2(Module):
    _input_ports = [('S', Float),
                    ('T', Float)]
    _output_ports = [('res', Float)]
    def compute(self):
        s = self.getInputFromPort("S")
        t = self.getInputFromPort("T")
        res = seawater.satN2(s, t)
        self.setResult('res', res)

_modules = [SaturationN2,
```

VisTrails Components
VisTrails Maya Plug-in

Provenance Enabling 3rd-Party Tools
Acknowledgments

• Python Community
• VisTrails Users & Developers
• VisTrails Team
  - Dr. Cláudio Silva
  - Dr. Juliana Freire

Questions