Provenance in Scientific Workflows: Challenges & Opportunities for Biodiversity Informatics

Bertram Ludäscher¹, Paolo Missier², David Lowery⁵, Tianhong Song¹, Maureen Kelly³, Sven Köhler¹, James Macklin⁵, Paul Morris³⁴, Robert Morris³, James Hanken⁴

¹ Dept. of Computer Science, UC Davis
² Computing Science, Newcastle University
³ Harvard University Herbaria
⁴ Museum of Comparative Zoology, Harvard
⁵ Agriculture and Agri-Food, Canada
Outline

• Scientific Workflows
  – Automation, Scalability, Abstraction, Provenance (ASAP)

• Kurator: Curation Workflows in Kepler (Filtered-Push)
  – Applying workflows to biodiversity data collections

• Provenance in Scientific Workflows (DataONE)
  – OPM to PROV to D-PROV

• Conclusions & Outlook
Outline

- Scientific Workflows
  - Automation, Scalability, Abstraction, Provenance (ASAP)

- Kurator: Curation Workflows in Kepler (Filtered-Push)

- Provenance in Scientific Workflows (DataONE)

- Conclusions & Outlook
Scientific workflows are similar to traditional scripts, in that they are used to automate computational pipelines e.g. data analysis steps.

Sci-wfs can be more scalable (e.g. may support pipeline and task parallelism etc)

They are also knowledge artifacts in their own right as they are often more abstract and thus easier to understand that complex programs or scripts.

Last not least, sci-wfs often support the recording of provenance metadata, specifically data lineage and processing history.

As a result, sci-wf can make computational experiments more transparent and easier to understand and reproduce.
Here you see an example of a bioinformatics workflow, called “Motif Catcher”.
On the left, the overall workflow is shown, while on the right, example data products used and produced in the workflow are depicted.
The workflow analyses sequence data, finds motifs, and then generates phylogenetic trees.
The initial workflow was implemented in Matlab.
To overcome some scalability issues, the workflow was reimplemented in Kepler, using the Map–Reduce extension available for Kepler.
While the previous slide showed a concept drawing of the workflow, here we are see an executable Kepler workflow. The map-reduce functions are defined as subworkflows. The overall workflow is still easy to recognize as it resembles the earlier conceptual workflows.
Outline

• Scientific Workflows
  – Automation, Scalability, Abstraction, Provenance (ASAP)

• Kurator: Curation Workflows in Kepler (Filtered-Push)

• Provenance in Scientific Workflows (DataONE)

• Conclusions & Outlook
The Problem:
Data Quality in Biodiversity Data

- **Collections & occurrence data** is "all over the map"
  - … literally (off the map!)
- **Issues**, e.g.:
  - Lat/Long transposition, coordinate & projection issues
  - Data entry/creation, "fuzzy" data, naming issues, bit rot, data conversions and transformations, schema mappings, … (you name it!)
- **Filtered-Push Collaboration**

The problem!
Here you see an example curation workflow from the FP project. In the first phase of the workflow, automated services are used to validate and if possible repair records. Data that cannot be automatically curated is routed to human experts, who receive an email with a request to inspect certain online spreadsheets to review and if necessary curated data records. Towards the end of the workflow, the curated dataset can be inspected, along with a provenance graph, that shows what data has been changed, by whom, and how.
### New Curation Provenance View

#### Quality Control By Data
- Analysis: Specify query parameters, select a workflow, and run.
- Query Parameters:
  - Property: Institution Code
  - Value: "SG" (Add Property/Value)
  - Workflow: "UpdateProvenance" (Add)

#### Summary of analysis (default to the last run or you may select a previous result from the list below):

<table>
<thead>
<tr>
<th>Institution Code</th>
<th>Collection Code</th>
<th>Catalog Number</th>
<th>Family</th>
<th>Scientific Name</th>
<th>Scientific Name Authority</th>
<th>Latitude</th>
<th>Longitude</th>
<th>County</th>
<th>State/Province</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
</tbody>
</table>

**TDWG 2013**

Provenance in Scientific Work
Kuration Workflow: Features & Challenges

• **Opportunities:**
  – **Automatic** capture of provenance information
    • data lineage and processing history
    • can be visualized, browsed, and queried
  – **Semi-automatic**: “Human-in-the-loop”
    • Domain experts as workflow actors!

• **Challenges:**
  – Automatic parts: many service invocations ➞ slow! 😞
    ➞ alternative, parallel execution models 😌
    (MapReduce, Akka process networks, … )
  – Semi-automatic (expert) curation: very long-running!
    ➞ develop suitable execution models
  – Provenance standards new, evolving; few tools
    • OPM ➞ W3C PROV ➞ DataONE D-PROV
Outline

- Scientific Workflows
  - Automation, Scalability, Abstraction, Provenance (ASAP)

- Kurator: Curation Workflows in Kepler (Filtered-Push)

- Provenance in Scientific Workflows (DataONE)

- Conclusions & Outlook
Many scientific disciplines increasingly make use of provenance metadata to be more transparent and to facilitate reproducible science.
This is also true for biodiversity workflows and in particular for curation workflows. After executing a curation workflow with provenance recording set to ON, the user can employ a provenance browsing and querying tool to “rewind the tape” and step through the processing history of the data used and produced by the workflow.
Here we are looking more closely at a data item (highlighted in red).
In the left pane we see more info about the data item.
We can also see which actor invocation (green box) created the data
And which one consumed it.
The tools has VCR-like controls to go forward and backward in the execution history.
... looking at an invocation...
Here is a quick summary of how workflow provenance can be used:
- Provenance is a form of evidence that allows the user to find out how a data product was derived, and whether and if so how it might be “tainted” by other tainted data products.
Alice, a climate scientist, has developed a UV-CDAT Vistrails workflow to generate benchmark gpp data. Once she has verified that the workflow generates the desired data, she creates a reproducible software package with ReproZip that enables other scientists to execute the workflow without the need to install and configure the particular libraries she is using. In addition, she exports the provenance information of the workflow execution and customizes it through the ProvExplorer tool, in order to eliminate the information she regards as superfluous. She then creates a data package with the ReproZip file, the customized provenance, and metadata. The package metadata is uploaded to a DataONE member node and indexed by a coordinating node.

Bob, another climate scientist, is looking for benchmark data to validate the climate model he has developed. He searches the DataONE repository and find Alice’s data package. He executes the ReproZip package to generate the benchmark data, which is used as input in the workflow he has developed along with his own data. The workflow generates a map projection and a Taylor diagram that enables him to verify the similarity between the
Here we see a slightly more abstract, but also expanded version of a collaborative workflow:
Alice develops (1) and runs (2) a workflow. Bob develops (3) and runs (5) a workflow using a variant of Alice’s shared data (4). If provenance is made available appropriately, Charlie can observe the “virtual collaboration” (via the shared data use).

We can view the composition $W_C$ as a new, virtual workflow:

- **How do we ensure that Charlie gets a complete account of the history of $W_C$’s outputs?**
- **How do we ensure that Alice gets her due (partial) credit when Charlie uses Bob’s data $v$?**
  - Provenance traces $T_A$ and $T_B$ will be critical
  - Need to compose them to obtain $T_C$
As part of the DataONE summer internship program, we have developed a couple of prototype tools for provenance management. This summer, we have prototyped a simple provenance repository, based on the Neo4j graph database.
Here is a cartoon overview of the prototype and some questions that we tried to answer. Note how different workflow systems (e.g. Taverna, Vistrails, Kepler etc) have different provenance importers. This summer, we only developed and used a Vistrails to D-PROV importer.
Let’s take a closer look at the provenance model, here OPM, the precursor of the W3C PROV standard. The key relations are “used” and “was generated by”. The former links a process P to its input data artifacts A. The latter links a data artifact A to its unique producer process P.

Edge labels are in past tense to express that these are used to describe prior executions.
Both OPM, and its W3C successor PROV provide some very basic information to capture “used” and “was-generated-by” information.

In our extended provenance model D-PROV (for DataONE PROV), we not only include this so-called trace-land information, but we also allow scientists to link workflow traces to the actual workflows that produced them.

In this way, using D-PROV, we can ask powerful queries that span trace-land, workflow-land (the wf specification), and more. The different colors indicate different kinds of provenance data.
Here are some more aspects of D-PROV that we need to add to the model

Aspects to cover with D-PROV

- Workflow modeling and structure (WF)
- Composition and subworkflows (SW)
- Workflow execution and traces (TR)
- Data representation and structure (DS)
- Workflow evolution (WE)
- Temporal data and constraints (TE)
- User and execution context information (CX)
Outline

- Scientific Workflows
  - Automation, Scalability, Abstraction, Provenance (ASAP)

- Kurator: Curation Workflows in Kepler (Filtered-Push)

- Provenance in Scientific Workflows (DataONE)

- Conclusions & Outlook
Summary & Next Steps

- Workflow technology yields new **opportunities** for (semi-)automatic data curation (FPush “Kurator” workflows)
- Prototype results are promising ...
- ... but challenges remain:
  - Develop execution models for large-scale, long-running workflows
  - Develop interoperable provenance model (DataONE/ProvWG):
    - D-PROV, based on W3C PROV, adds “workflow land” and “data land”

- **Call for Action:**
  - Bring together community of curation workflow enthusiasts
  - Develop important provenance use cases for curation workflows