Taverna, Biocatalogue, myExperiment, and the provenance of it all: forward-looking while looking back

Dr. Paolo Missier, Prof. Carole Goble
Information Management Group
School of Computer Science, University of Manchester, UK

with additional material kindly provided by:
Prof. Dave DeRoure, Prof. Luc Moreau, Univ. of Southampton, UK
Andrea Wiggins, Syracuse University, NY
Part I: models and technology for e-science

1. Addressing the needs of the e-scientist:
   – Workflow as a model of experimental science
     - Taverna
     - Services as building blocks
     - Biocatalogue

2. Scaling up along the social dimension:
   - towards open, collaborative science
   - myExperiment

Part II: Explaining and Preserving experimental outcomes

- Data provenance support in Taverna
- provenance for open science: the OPM vision
What is the myGrid Project?

- UK e-Science pilot project since 2001.
- Centred at Manchester, Southampton and the EMBL-EBI
- Part of Open Middleware Infrastructure Institute UK [http://www.omii.ac.uk](http://www.omii.ac.uk).
- Mixture of developers, bioinformaticians and researchers
- An alliance of contributing projects and partners
- Open source development and content LGPL or BSD
- Infrastructure
- We don’t own any resources (apart from catalogues)
- Or a Grid.
Workflows: *E. Science laboris*

- Pipeline processing
- Automated processing
- Repetitive and mundane boring stuff made easier, reliable and adaptable.
- Shield interoperability horror
- Trackable results
- Agile software development
- Big science, small science & collaborative science
Workflow as data integrator

QTL genomic regions

genes in QTL

metabolic pathways (KEGG)
Workflow as data integrator

QTL genomic regions

genes in QTL

metabolic pathways (KEGG)
List-structured KEGG gene ids:

\[
\text{[[ mmu:26416 ], [ mmu:328788 ]]}\]

[ path:mmu04010 MAPK signaling,
  path:mmu04370 VEGF signaling ]

[ [ path:mmu04210 Apoptosis, path:mmu04010 MAPK signaling, ...],
  [ path:mmu04010 MAPK signaling, path:mmu04620 Toll-like receptor, ...] ]
List-structured KEGG gene ids:

\[
[ [ \text{mmu:26416} ], \text{mmu:328788} ]
\]

\[
[ \text{path:mmu04010 MAPK signaling}, \text{path:mmu04370 VEGF signaling} ]
\]

\[
[ [ \text{path:mmu04210 Apoptosis}, \text{path:mmu04010 MAPK signaling} ], \text{path:mmu04620 Toll-like receptor} ]
\]
List-structured KEGG gene ids:

- \([ [ \text{mmu:26416} ], [ \text{mmu:328788} ] ]\)

- \([ \text{path:mmu04010 MAPK signaling, path:mmu04370 VEGF signaling} \] \)

- \([ [ \text{path:mmu04210 Apoptosis, path:mmu04010 MAPK signaling, ...}], [ \text{path:mmu04010 MAPK signaling, path:mmu04620 Toll-like receptor, ...} ] \)
What do Scientists use Taverna for?

**Systems biology model building**
Proteomics
Sequence analysis
Protein structure prediction
Gene/protein annotation
Microarray data analysis
QTL studies
QSAR studies
Medical image analysis
Public Health care epidemiology
Heart model simulations
High throughput screening
Phenotypical studies
Phylogeny
  - Statistical analysis
  - Text mining
Astronomy, Music, Meteorology

**Netherlands Bioinformatics Centre**
Genome Canada Bioinformatics Platform
BioMOBY
US FLOSS social science program
RENCI
SysMO Consortium
French SIGENAE farm animals project
ThaiGrid
CARMEN Neuroscience project
SPINE consortium
EU Enfin, EMBRACE, BioSapian, Casimir
EU SysMO Consortium
NERC Centre for Ecology and Hydrology
Bergen Centre for Computational Biology
Max-Planck institute for Plant Breeding Research
Genoa Cancer Research Centre
AstroGrid

30 USA academic and research institutions
Genotype

200

Metabolic pathways

Phenotype

Literature

[Paul Fisher]
KnowARC integrated with Taverna" application prototype to use Taverna as direct interface to Grid resources running ARC.

http://www.knowarc.eu

Open source grid software infrastructure aimed at enabling multi-institutional data sharing and analysis. Underpins caBIG. Taverna links together caGrid resources.

http://cagrid.org/

Europe’s leading grid computing project, Piloted Taverna over EGEE gLite services

http://www.eu-egee.org/
- Service-oriented applications
  - Applications components of workflows
  - Compose applications into workflows
  - Incorporate workflows into applications

- Service-oriented Grid Infrastructure
  - Provision physical resources to support application workflows
  - Coordinate resources through workflows

---

myGrid

Composition Incorporation

Invocation

Provisioning Workflows

Users

Workflows

Appln Service

Appln Service

Workflows

[Foster 2005]
Who else is in this space?

Trident

Kepler

Taverna

Triana

BPEL

BioExtract
Workflow-based experimentation lifecycle

- **Run**
- **Collect and query provenance metadata**
- **Analyse Results**
- **Publish**
- **Develop**

**Tools**
- **Taverna 2**
- **BioCatalogue**
- **myexperiment**
Workflow-based experimentation lifecycle

- Run
- Analyse Results
- Publish
- Develop

Collect and query provenance metadata

Taverna 2

BioCatalogue

my experiment
Graphical Workbench For Professionals

Plug-in architecture
Nested Workflows
Drag and Drop
Wiring together

Rapidly incorporate new service without coding.
Not restricted to predetermined services
Access to local and remote resources and analysis tools
3500+ service operations available when start up
Services Mutability

implications for sustainability, accountability and reproducability

- Reliability and robustness of workflows depends on the reliability and robustness of the components
- In house service support
- Services in constant and (silent) change.
- Versioning.
- Workflow Decay
- Monitoring and Repair of wrappers, shims and service substitutions.
http://www.biocatalogue.org
http://beta.biocatalogue.org

Professor Carole Goble
University of Manchester, UK
Director myGrid Consortium

28 April 2009, Boston MA

Data curation + process curation=data integration + science
Briefings in Bioinformatics, doi:10.1093/bib/bbn034 (Dec., 2008)
Carole Goble, Robert Stevens, Duncan Hull, Katy Wolstencroft and Rodrigo Lopez
The short story

• Public, Curated Catalogue of Life Science Web Services
• Register, Find, Curate Web Services
• Community-sourced annotation, expert oversee
• Open content

• Open platform with open REST interfaces
• Web 2.0 site and development.
• Open source code base.

• Started June 2008. In first beta phase.
• Launched June 2009 at ISMB.
• www.biocatalogue.org
The short story

- Public, Curated Catalogue of Life Science Web Services
- Register, Find, Curate Web Services
- Community-sourced annotation, expert oversee
- Open content

- Open platform with open REST interfaces
- Web 2.0 site and development.
- Open source code base.

- Launched June 2009 at ISMB.
- www.biocatalogue.org
The short story

• Public, Curated Catalogue of Life Science Web Services
• Register, Find, Curate Web Services
• Community-sourced annotation, expert oversee
• Open content

• Open platform with open REST interfaces
• Web 2.0 site and development.
• Open source code base.

• Started June 2008. In first beta phase.
• Launched June 2009 at ISMB.
• www.biocatalogue.org
A virtual circle

- curation involves substantial human effort
  why would it happen at all?
Curation

- Just enough just in time
- Universal annotation scheme
- Mixed: Free text, Tags, controlled vocabns, community ontologies
- Community sourced tags, comments, recommendations
- Expert curation ontology-based annotation. myGrid OWL Ontology
- Automated WSDL ripping and analytics
- Automated monitoring & testing
- Partner feeds (e.g. myExperiment)
- Update feeds to users

Today: 14902 annotations (provider, user, registries)
KEGG: 1433 annotations
Service monitoring

The EMBRACE Service Registry is a collection of life–science web services with built-in service testing.

This site is a prelude to the internationally supported BioCatalogue system that will collect, store, validate, and make available web–services in the biosciences. This registry is mainly meant for the EU projects EMBRACE, BioSapiens and ENFIN, but other users are welcome too. As a potential web service user, you can search or browse the registry for services that match your needs. Furthermore, each entry includes live test data, showing
Scaling up along the social dimension

Crossing the boundaries of individual investigation

Develop Run Analyze Publish
Scaling up along the social dimension

Crossing the boundaries of individual investigation
Scaling up along the social dimension

Crossing the boundaries of individual investigation
Scaling up along the social dimension

Crossing the boundaries of individual investigation

Scientific Collaboration Requirements

- Shared goals
  - Establishes focus of research

- Shared research resources
  - Both social and artifactual
  - Social aspects include training and community socialization

Source: Andrea Wiggins, talk given at the School of Computer Science, University of Manchester, UK, June 18th, 2009
Historical Research Artifacts

• Letters, Books, Journals, Lectures

• Also technologies: methods, instrumentation

• Sharing?

  • Recordkeeping is not always a researcher’s main priority

  • Without records, there’s not much to share except the research outputs

Source: Andrea Wiggins, talk given at the School of Computer Science, University of Manchester, UK, June 18th, 2009
Today’s Research Artifacts

• Large scale datasets, scripts, software, workflows, papers, images, video, audio, annotations, ephemera, web sites...

• “Research objects” - bundling all the pieces together

• Hybrids of boundary objects and touchstones

• Technologies -> scientific revolution!

• Open science

Source: Andrea Wiggins, talk given at the School of Computer Science, University of Manchester, UK, June 18th, 2009
The Selfish (or Self-interested) Scientist

“A biologist would rather share their toothbrush than their gene name”

Mike Ashburner and others
Professor Genetics,
University of Cambridge, UK

“Data mining: my data’s mine and your data’s mine”
The potential for collaboration

• **What:**
  – processes: “materials and methods” → workflows
  – data: unlikely, and certainly not until published
  – metadata (annotations, provenance traces...): ??

• **When:**
  – for contributors: part of publication process
    • some publishers demand public data and repeatable experiments
  – for consumers: reuse as part of experiment design

• **Where and how:**
  – a meeting point for a virtual community
  – Web 2.0 style of interaction
  – voluntary, incentive-based contributions
Asymmetric vs symmetric sharing

Traditional sharing is asymmetric:

• Producer-consumer:
  – from service providers to workflow designers
  – Biocatalogue

Open science is symmetric:

• Peer-based
  – sharing of workflows as complex processes
  – myExperiment

myGrid combines both paradigms:

• Service space “closed under composition”:
  – workflows are compositions of services
  – ... and they are services themselves

• Scientists become providers
  – of conceptual process models
  – and of executable services, as well!
Design-time reuse: Composition from existing workflows

Run

Collect and query provenance metadata

Publish

Compare results across versions

Analyse Results

Foster virtual scientific communities

Runtime reuse: Workflows as services

Provenance exchange and interoperability - the OPM experiment
Publishing for collaboration

Run

Collect and query provenance metadata

Analyse Results

myexperiment

Design-time reuse: Composition from existing workflows

Runtime reuse: Workflows as services

provenance exchange and interoperability the OPM experiment

compare results across versions

foster virtual scientific communities

ESIP meeting, Santa Barbara, CA, July 2009 - P. Missier
Collaboration in the workflow space

Collaboration in the workflow space

Develop → Run → Analyze → Publish

myexperiment
Competitive advantage.
Academic vanity.
Adoption.
Reputation.

Scrutiny.
Being scooped.
Misinterpretation.
Reputation.
• Getting author to take credit!
• Creating a culture of attribution.
• Attribution and credit chains.
• Licensing and rights protection
Incentive and reputation

- Strong sense of persistent identity.
- Building reputation and boasting opportunities.
- Cult of the individual.
- High visibility to the participant and the community.
- Downloads & Views.
- Instrumentation and automated analysis.
- Feedback.
- Liability policy.
Reuse, Recycling, Repurposing
Cross-fertilization

- Paul writes workflows for identifying biological pathways implicated in resistance to Trypanosomiasis in cattle.

- Paul meets Jo. Jo is investigating Whipworm in mouse.

- Jo reuses one of Paul's workflow without change.

- Jo identifies the biological pathways involved in sex dependence in the mouse model, believed to be involved in the ability of mice to expel the parasite.

- Previously a manual two year study by Jo had failed to do this.
Socially share, discover and reuse workflows and other methods.

Cooperative bazaar.

Sunday 10th May: 1748 registered users, 143 groups, 669 workflows, 197 files, 52 packs 56 different countries. Top 4: UK, US, The Netherlands, Germany

www.myexperiment.org
Title: Escherichia coli: From cDNA Microarray Raw Data to Pathways and Published Abstracts

Type: Taverna 1

Download Scalable Diagram (SVG)
Self-Curation by Contributors

Social Curation by the Crowd

Curation by Experts

Automated Curation

myexperiment

BioCatalogue
### Taverna 1.7.1 starter pack

**Created:** 17/07/08 @ 21:06:12  |  **Last updated:** 20/07/08 @ 15:46:51

<table>
<thead>
<tr>
<th>Everything to get started with Taverna 1.7.1</th>
</tr>
</thead>
</table>

| 16 items in this pack |

**Comments:** 0  |  **Viewed internally:** 215 times  |  **Downloaded internally:** 47 times

**Tags:**
- example
- introduction
- tutorial
Towards Genotype-Phenotype Correlations

Created: 08/04/09 @ 13:14:54 | Last updated: 08/04/09 @ 13:16:23

It is increasingly common to combine Microarray and Quantitative Trait Loci data to aid the search for candidate genes responsible for phenotypic variation. Workflows provide a means of systematically processing these large datasets and also represent a framework for the re-use and the explicit declaration of experimental methods. In this pack is a paper which describes the issues facing the manual analysis of microarray and QTL data for the discovery of candidate genes underlying complex phe...

19 items in this pack

Comments: 0 | Viewed internally: 4 times | Downloaded internally: 0 times

Tags:
affymetrix | african trypanosomiasis | cattle | data-driven | disease | entrez | genotype | Kegg Pathways | KeggID | link-integration | microarray | mouse | pathway | pathway-driven | phenotype | sleeping sickness | swissprot | uniprot | web services
Towards Genotype-Phenotype Correlations

It is increasingly common to combine Microarray and Quantitative Trait Loci data to aid the search for candidate genes responsible for phenotypic variation. Workflows provide a means of systematically processing these data sets, but data management and annotation are often manual and ad hoc processes.

myExperiment paper for Concurrency Practice and Experience

eScience 2008 Special Issue

Towards Genotype-Phenotype Correlations

Created: 08/04/09 @ 13:14:54 | Last updated: 08/04/09 @ 13:16:23

It is increasingly common to combine Microarray and Quantitative Trait Loci data to aid the search for candidate genes responsible for phenotypic variation. Workflows provide a means of systematically processing data, declaring annotations, and manually managing results.

myExperiment paper for Concurrency Practice and Experience

eScience 2008 Special Issue

Created: 10/04/09 @ 13:41:56 | Last updated: 14/04/09 @ 08:34:12


Workflow discovery benchmarks

Created: 12/07/08 @ 11:20:23 | Last updated: 09/10/08 @ 16:29:50

This pack contains benchmarks that measure how bioinformaticians discover Taverna workflows. Several subpacks are available: Collection of workflows by Paul Fischer, used in benchmarks PR2 and CA2. Collection of workflows by Peter Li, used in benchmarks PR2 and CA2.

7 items in this pack

Comments: 0 | Viewed internally: 52 times | Downloaded internally: 2 times
Tags:
benchmarks
Collaboration in the workflow space

- Develop
- Run
- Analyze
- Publish

Technical implications of open science

- **Process interoperability**
  - SOA principles: runtime interoperability
  - but, still no common workflow model after all!

- **Data interoperability**
  - Traditional heterogeneity / integration issues
  - Dataspaces
  - LinkedData
  - ...

- **Aggregation: creating logical units**
  - process + inputs + outputs + provenance traces + ...
  - Research Objects

- **Provenance interoperability**
Technical implications of open science

• Process interoperability
  – SOA principles: runtime interoperability
  – but, still no common workflow model after all!

• Data interoperability
  – Traditional heterogeneity / integration issues
  – Dataspaces
  – LinkedData
  – ...

• Aggregation: creating logical units
  – process + inputs + outputs + provenance traces + ...
  – Research Objects

• Provenance interoperability
Provenance of data

• “It would include details of the processes that produced electronic data as far back as the beginning of time or at least the epoch of provenance awareness.”

Analysis of process results

• Causal relations:
  - which pathway sets come from which gene sets?
  - which processes contributed to producing this image?
  - which process(es) caused this data to be incorrect?
  - which data caused this process to fail?

• Process and data analytics:
  - show me the variations in output in relation to an input parameter sweep (multiple process runs)
  - how often has my favourite service been executed?
    • on what inputs?
  - who produced this data?
  - how often does this pathway turn up when the input genes range over a certain set S?
List-structured KEGG gene ids:

```python
[ [ mmu:26416 ], [ mmu:328788 ] ]
```

Workflow Inputs:
- `list_of_genIDList`

Workflow Outputs:
- `paths_per_gene`
- `commonPathways`

Example:

- `get_pathways_by_genes`
- `getCommonPathways`
- `getPathwayDescriptions`
- `getPathwayDescriptions2`
- `get_inputlist`
- `Flatten_List`
- `outputlist`

Pathways and geneIDs:

- `path:mmu04010 MAPK signaling`
- `path:mmu04370 VEGF signaling`
- `path:mmu04010 MAPK signaling`
- `path:mmu04620 Toll-like receptor`
- `path:mmu04210 Apoptosis`

- `geneIDs`:
  - `mmu:26416`
  - `mmu:328788`

- `pathways`:
  - Connections between geneIDs and pathways.
Example: inverse data associations

List-structured KEGG gene ids:

```python
[ [ mmu:26416 ], [ mmu:328788 ] ]
```

GeneIDs:

- mmu:26416
- mmu:328788

Pathways:

- path:mmu04010 MAPK signaling
- path:mmu04370 VEGF signaling
- path:mmu04010 MAPK signaling
- path:mmu04620 Toll-like receptor

```python
[ [ path:mmu04210 Apoptosis, path:mmu04010 MAPK signaling, ...], [ path:mmu04010 MAPK signaling, path:mmu04620 Toll-like receptor, ...] ]
```
Example: inverse data associations

List-structured KEGG gene ids: 
[[ mmu:26416 ], [ mmu:328788 ]]

[ path:mmu04010 MAPK signaling, path:mmu04370 VEGF signaling ]

[ [ path:mmu04210 Apoptosis, path:mmu04010 MAPK signaling, ...], [ path:mmu04010 MAPK signaling, path:mmu04620 Toll-like receptor, ... ] ]
Example: inverse data associations

List-structured KEGG gene ids:

```
[[ mmu:26416 ],
 [ mmu:32878 ]]
```

GeneIDs:

- path:mmu04010 MAPK signaling
- path:mmu04370 VEGF signaling

Pathways:

- Apoptosis
- MAPK signaling
- Toll-like receptor

Workflow Inputs:

- list_of_geneIDList
- genes_id_list
- get_pathways_by_genes
- attachmentList
- inputlist
- Flatten_List
- outputlist

Workflow Outputs:

- paths_per_gene
- commonPathways

```
[[ path:mmu04210 Apoptosis, path:mmu04010 MAPK signaling, ...],
 [ path:mmu04010 MAPK signaling , path:mmu04620 Toll-like receptor, .. ]
```
• Taverna type system: strings + nested lists
  – “cat”, [“cat”, “dog”], [ [“cat”, “dog”], [“large”, “small”] ]

• Taverna dataflow model: data-driven execution
  • services activate when input is ready

• Workflow provenance: a detailed trace of workflow execution
  – which services were executed
  – when
  – inputs used, outputs produced
Taverna + provenance

• Taverna type system: strings + nested lists
  – “cat”, [“cat”, “dog”], [ [“cat”, “dog”], [“large”, “small”] ]

• Taverna dataflow model: data-driven execution
  • services activate when input is ready

• Workflow provenance: a detailed trace of workflow execution
  – which services were executed
  – when
  – inputs used, outputs produced

Taverna dataflow model + provenance traces can be a powerful combination
Focus is on the data: the **observable outcomes of a process**

<table>
<thead>
<tr>
<th>raw provenance metadata</th>
<th>provenance metadata + interpretation framework</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>design</strong></td>
<td>• service annotations:</td>
</tr>
<tr>
<td></td>
<td>• ex. get_pathways_by_genes</td>
</tr>
<tr>
<td></td>
<td>• who created /edited: attribution</td>
</tr>
<tr>
<td></td>
<td>• why: purpose, intent</td>
</tr>
<tr>
<td>• process structure (workflow graph)</td>
<td>• process events:</td>
</tr>
<tr>
<td>• history of process composition - reuse</td>
<td>- service invocation</td>
</tr>
<tr>
<td>• process versions</td>
<td>- data production / consumption</td>
</tr>
<tr>
<td></td>
<td>- causal dependency graphs</td>
</tr>
<tr>
<td></td>
<td>ex.:</td>
</tr>
<tr>
<td></td>
<td>- list_of_geneIDList = [ a, b, c]</td>
</tr>
<tr>
<td></td>
<td>- paths_per_gene = [ [d,e,f], [g,h,j]]</td>
</tr>
<tr>
<td></td>
<td>- ... in run #32</td>
</tr>
<tr>
<td></td>
<td>- data annotations,</td>
</tr>
<tr>
<td></td>
<td>results interpretation in terms of conceptual data model:</td>
</tr>
<tr>
<td></td>
<td>set of pathways → gene sets</td>
</tr>
</tbody>
</table>

- **Focus is on the data:** the observable outcomes of a process

- **raw provenance metadata**

- **provenance metadata + interpretation framework**

- **design**
  - process structure (workflow graph)
  - history of process composition - reuse
  - process versions

- **execution**
  - process events:
    - service invocation
    - data production / consumption
    - causal dependency graphs
      ex.:
      - list_of_geneIDList = [a, b, c]
      - paths_per_gene = [ [d,e,f], [g,h,j]]
      - ... in run #32
...and their uses and associated challenges

<table>
<thead>
<tr>
<th>raw provenance metadata</th>
<th>provenance metadata + interpretation framework</th>
</tr>
</thead>
</table>
| **design**              | • exploiting semantic properties of the process structure to improve provenance exploitation  
                          | • exploring process space across versions and structural similarities  
                          | • graph matching  
                          | • semantic-based search of process space |
| **execution**           | - enabling partial re-runs of resource-intensive workflows  
                          | - storing very large provenance traces that accumulate over time  
                          | - efficient query over large traces  
                          | - presentation of query answers  
                          | - semantic-based query answering over annotated traces |
...and their uses and associated challenges

<table>
<thead>
<tr>
<th>raw provenance metadata</th>
<th>provenance metadata + interpretation framework</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>design</strong></td>
<td></td>
</tr>
<tr>
<td>• exploiting semantic properties of the process structure to improve provenance exploitation</td>
<td>• semantic-based search of process space</td>
</tr>
<tr>
<td>• exploring process space across versions and structural similarities</td>
<td></td>
</tr>
<tr>
<td>• graph matching</td>
<td></td>
</tr>
<tr>
<td><strong>execution</strong></td>
<td></td>
</tr>
<tr>
<td>- enabling partial re-runs of resource-intensive workflows</td>
<td>- semantic-based query answering over annotated traces</td>
</tr>
<tr>
<td>- storing very large provenance traces that accumulate over time</td>
<td></td>
</tr>
<tr>
<td>- efficient query over large traces</td>
<td></td>
</tr>
<tr>
<td>- presentation of query answers</td>
<td></td>
</tr>
</tbody>
</table>

fully implemented in Taverna 2

to be released in Sept. 2009
...and their uses and associated challenges

<table>
<thead>
<tr>
<th>raw provenance metadata</th>
<th>provenance metadata + interpretation framework</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>design</strong></td>
<td></td>
</tr>
<tr>
<td>• exploiting semantic properties of the process structure to improve provenance exploitation</td>
<td>• semantic-based search of process space</td>
</tr>
<tr>
<td>• exploring process space across versions and structural similarities</td>
<td></td>
</tr>
<tr>
<td>• graph matching</td>
<td></td>
</tr>
<tr>
<td><strong>execution</strong></td>
<td></td>
</tr>
<tr>
<td>- enabling partial re-runs of resource-intensive workflows</td>
<td>- semantic-based query answering over annotated traces</td>
</tr>
<tr>
<td>- storing very large provenance traces that accumulate over time</td>
<td></td>
</tr>
<tr>
<td>- efficient query over large traces</td>
<td></td>
</tr>
<tr>
<td>- presentation of query answers</td>
<td></td>
</tr>
</tbody>
</table>

fully implemented in Taverna 2

to be released in Sept. 2009
Querying provenance traces

- Lineage queries involve traversing a *provenance graph* from bottom to top.
Naive provenance trace queries

- In most approaches, the originating process are not used for querying
- consequence: query requires provenance graph traversal
  - large traces $\rightarrow$ computationally complex
  - view materialization used in practice to get around the computational complexity

![Diagram of workflow processes](image)

Fig. 1. Provenance graphs

• Users are rarely interested in the complete provenance graph
  – noisy, possibly large, difficult to navigate

This results in a more efficient lineage query algorithm that scales to large provenance graphs
Querying provenance graphs in Taverna

- Users are rarely interested in the complete provenance graph
  - noisy, possibly large, difficult to navigate

This results in a more efficient lineage query algorithm that scales to large provenance graphs
Provenance management architecture

- Taverna runtime
- workflow inputs
- provenance events
  - input arrived
  - service invoked
  - output produced
- Provenance capture component
- workflow results
- Results analysis
  - Results browser
  - Provenance browser
- Lineage query processor
- Provenance DB
  - relational data model
OPM: the Open Provenance Model
Provenance Across Applications

Adapted from Luc Moreau’s slides: “The Open Provenance Model” (Univ. of Southampton, UK), 2009
Provenance Across Applications

Adapted from Luc Moreau’s slides: “The Open Provenance Model” (Univ. of Southampton, UK), 2009
Illustration

- Process “used” artifacts and “generated” artifact
- Edge “roles” indicate the function of the artifact with respect to the process (akin to function parameters)
- Edges and nodes can be typed

**Causation chain:**
- P was caused by A1 and A2
- A3 and A4 were caused by P
- Does it mean that A3 and A4 were caused by A1 and A2?

From Luc Moreau’s slides set: “The Open Provenance Model” (Univ. of Southampton, UK), 2009
Integrated OPM generation in Taverna
Integrated OPM generation in Taverna
the answer to any TP query can be viewed as an OPM graph
encoded as RDF/XML using the Tupelo provenance API (NCSA)
More information on OPM

• The OPM wiki:
  – http://twiki.ipaw.info/bin/view/OPM/
    • open to discussions and contributions
    • please read the governance doc

• The 3rd provenance challenge:
  – produce and export OPM graphs
    • interoperable XML and RDF serializations
  – import and query third party graphs

  – http://twiki.ipaw.info/bin/view/Challenge/ThirdProvenanceChallenge
  – The University of Manchester’s contribution to the challenge:
    • http://twiki.ipaw.info/bin/view/Challenge/UoM
  – latest meeting held in June, 2009 (Amsterdam)
answering user questions effectively
(using provenance + semantics infrastructure)
- has a similar investigation been undertaken before? when, by whom?
  what was the outcome?
- have alternative services being used? to what effect?
- what have been the users’ decisions, and why?

Enabling collaborative science:
- provide users with recommendations on the next steps in their session,
  based on analysis of their past actions;
- cluster users within a group based on their common interests, observed
  through the choices they make during the sessions
- promote socal/scientific networking
  • a “blog the lab” flavour
From knowledge capture to exploitation

➡ answering user questions effectively
  (using provenance + semantics infrastructure)
  – has a similar investigation been undertaken before? when, by whom? what was the outcome?
  – have alternative services being used? to what effect?
  – what have been the users’ decisions, and why?

➡ Enabling collaborative science:
  – provide users with recommendations on the next steps in their session, based on analysis of their past actions;
  – cluster users within a group based on their common interests, observed through the choices they make during the sessions
  – promote social/scientific networking
    • a “blog the lab” flavour

(Semantic) provenance analytics
The Provenir ontology

- Upper ontology with for domain-specific extensions
- OWL, designed for reasoning and RDF queries

Satya S. Sahoo, Roger S. Barga, Jonathan Goldstein, Amit P. Sheth, *Where did you come from...Where did you go?* An Algebra and RDF Query Engine for Provenance, TR-2009-03, Kno.e.sis Center, CSE Dept., Wright State University, Dayton, OH, March, 2009
### Upcoming events

| **SWPM 2009:**  
The First International Workshop on  
the Role of Semantic Web in Provenance Management |
|-----------------------------------------------|
| Co-located with ISWC'09, October 25/26 2009, Washington D.C., USA  
Submission Deadline: Friday, July 31, 2009 |

| **Special issue of Future Generation Computer Systems Journal** (FGCS)  
on the third provenance challenge  
(to be announced) |
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>expected deadline: <strong>Dec., 2009</strong></td>
</tr>
</tbody>
</table>
Summary: Support for collaborative science

Provenance:
- automated metadata collection and processing
1. data management angle: “logs with a proper data model”
   - storage and query issues
   - interoperability

2. social angle: attribution chain for experimental artifacts
   - processes, data, annotations
   - myExperiment packs → Research Objects
Selected literature on provenance