Provenance and the Semantic Web

Jiao Tao, Zhenning Shangguan, Rui Huang, James Michaelis

CSCI 6965 – Advanced Semantic Technologies (Class 8)

3 / 17 / 2009
Our Plan

• Define what Provenance is
  – Current application areas
  – How it is represented/structured

• Address how Semantic Web backed provenance is useful and can be implemented
  – Current research

• Concluding Remarks
What is Provenance?

• Can denote the following things about something:
  – Origin
  – History
  – Ownership
  – Location
Applications of Provenance

• Used in many fields (antique handling, library science, etc.)

• Relevance to natural sciences:
  – Activity auditing
  – Reproducing prior findings
  – Data categorization
  – Determining data access rights
  – Data exploration programs
Requirements for Provenance

• Adequately structured domain data
  – Produced in reasonable quantities in many complex systems

• Model for handling representations
  – A fair number exist (for instance PML, OPM)
  – However, many of them have common structural features
Provenance Models

• What do they look like?
  – Varies, but for our purposes they tend to have the following definitions:
  – Artifact: Some section of data
  – Agent: An entity responsible for manipulating data
  – Activity: The function performed by an Agent on an Artifact
  – These three can be grouped into a data handling Event.
How is Semantic Web Relevant?

• On use for Provenance representation:
  – “It is easy to represent annotations and connections between processes and files. Furthermore, the semantics of RDF and OWL allow us to define terms such that a reasoner can infer connections among files and processes. Given a specific file, all of the direct and indirect connections it has to other files and processes can be discovered with a simple query.”

• In A Semantic Web Approach to the Provenance Challenge. Golbeck and Hendler, 2006
Examples of Current Research

• Use of Semantic Web technologies for representing/handling provenance.
  – *A Semantic Web Approach to the Provenance Challenge*, Golbeck, J. Hendler, J.

• Generating high-quality semantic provenance metadata
  – *Semantic Provenance for eScience: Managing the Deluge of Scientific Data*

• Facilitating information visualization and access policies
  – Provenance Explorer - Tailored Provenance Views Using Semantic Inferencing, ISWC2006
A Semantic Web Approach to the Provenance Challenge

Concurrency and Computation: Practice & Experience 2008

Jennifer Golbeck, James Hendler
University of Maryland, College Park
Motivation: First Provenance Challenge

http://twiki.ipaw.info/bin/view/Challenge/FirstProvenanceChallenge
SW Approach to the Challenge

- Execute challenge workflow through web services
- Represent provenance schema with ontologies
- Store provenance data in triple stores
- Infer new knowledge with OWL reasoning and rules
- Implement Challenge Queries with SPARQL

http://provenance.mindswap.org
Web Service-based Workflow Execution

- Each process is encapsulated in a web service.
- The service produces:
  - Output files.
  - OWL file describing the provenance information for its execution and for each output file.
- Each input/output/ontology file has an URI.
- WES (Workflow Execution Service) executes the whole workflow:
  - Uses initial input files as inputs
  - Passes correct inputs to the services representing the processes, receives the outputs, and passes to next service, ...
  - Finally, outputs the URIs of the final files, the three Atlas Graphics, and an OWL file describing its execution.
Provenance Ontology

• Service Execution
  – Input and Output files
  – Input parameters
  – Date and time
  – Connection to workflow execution
  – Service used
  – Stage

• File
  – File parents and file ancestors
  – Produced by Service (web service)
  – Service Execution Parents (specific execution)
  – Siblings
  – Date and Time information

• Workflow Execution
  – Service Executions
  – Times

Ontology: http://provenance.mindswap.org/provenance.owl
Some information encoded as rules, e.g. Service Execution Ancestry
Implementing Challenge Queries with SPARQL

• Query 1: find the process that led to Atlas X Graphic / everything that caused Atlas X Graphic to be as it is. This should tell us the new brain images from which the averaged atlas was generated, the warping performed etc.

• Solution:

```sparql
SELECT DISTINCT ?property ?value WHERE {
}
```

• NOTE:
  – The file that contains all of the data for the workflow is assigned the prefix “workflow”.
  – Atlas X Graphic: workflow:Graphic1155932184.62982
Implementing Challenge Queries with SPARQL

• Query 2: find the process that led to Atlas X Graphic, excluding everything prior to the averaging of images with softmean.

• Solution:

```sparql
PREFIX prov: <http://provenance.mindswap.org/provenance.owl#>
SELECT DISTINCT ?subject ?property ?object WHERE {
  workflow:Graphic1155932184.62982 ?prop1 ?subject
  FILTER (?prop1 = prov:hasServiceExecutionAncestor).
  ?file ?prop2 ?service
  FILTER (?prop2 = prov:hasServiceExecutionAncestor).
  ?service prov:serviceUsed prov:softmean.}
```

• NOTE: show only processes (ServiceExecutions) and their data where softmean is the service used by ServiceExecutionAncestor
Implementing Challenge Queries with SPARQL

• Query 9: a user has annotated some atlas graphics with key-value pair where the key is studyModality. Find all the graphical atlas sets that have metadata annotation studyModality with values speech, visual or audio, and return all other annotations to these files.

• Solution:

```sparql
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX prov: <http://provenance.mindswap.org/provenance.owl#>
SELECT DISTINCT ?graphic ?annotation WHERE {
    ?graphic rdf:type prov:Graphic.
    ?graphic prov:annotation ?oneOfThese
    FILTER ( ?oneOfThese = "studyModality=speech" ||
             ?oneOfThese = "studyModality=audio" ||
             ?oneOfThese = "studyModality=visual" ).
}
```
Semantic Provenance for eScience: Managing the Deluge of Scientific Data

Satya S. Sahoo, Amit Sheth, and Cory Henson
Wright State University
Outline

• Motivation
• Semantic Provenance
• Two Degrees of Separation
• Semantic Provenance Framework
• Implementation: Spade
Motivation

• Challenge
  – Scientists collaborate via worldwide infrastructure of computing and data resources
  – Huge amount of scientific data

• Goal
  – Computing solutions that use high-quality metadata to interpret and manage scientific data
Semantic Provenance

- One form of general metadata
- Refer to a formal knowledge model or an ontology that imposes a domain-specific provenance view on scientific data
- Consists of concepts, relationships, and constraints

\[
\text{Semantic Provenance} = \text{Provenance} \ + \ \text{Domain Knowledge} \ + \ \text{Underlying Ontology!}
\]
Two Degrees of Separation

• Problem
  – Traditional work-flow-centric approach of generating provenance information cannot answer queries with domain semantics
  – e.g., “Find proteins composed of peptides with N-glycosylation consensus sequence {*N[^P][S/T]*}”

• Solution
  – Create semantic provenance using specialized services
  – Refer to domain-specific provenance ontology during creation
  – Can be integrated into the workflow engine on demand
Semantic Provenance Framework

Semantic provenance annotation
For example:
Time  Space  Theme

Domain provenance ontologies
Agent (continuant)
For example, a sensor

Process (occurent)
For example, filtering or merging

Data (continuant)
For example, a temperature reading

Reasoning, query answering
Trust, security
Integration, visualization

Applications
Spade

• Features
  – Realization of the Semantic Provenance Framework in the glycoproteomics domain
  – Automate the scientific data analysis process as a workflow using Semantic Web services plus Taverna workflow engine
  – ProPreo proteomics provenance ontology
  – Semantic provenance creation services
Architecture

- User
- SPARQL query interface
- ProPreO ontology
- Semantic provenance information
- Semantic provenance annotation modules
- Semantic provenance module 1
- Semantic provenance module 2
- Semantic provenance module 6
- Oracle RDF store
- Mass spectrometry data-analysis protocol
  - raw2mzXML
  - mzXML2pkl
  - pkl2pSplit
  - Mascot database search
  - ProValt application
ProPreo ontology
Provenance Explorer – Customized Provenance Views Using Semantic Inferencing

In Proceedings of ISWC 2006

Kwok Cheung, Jane Hunter
The University of Queensland
Objective

Take the output from RDF-based workflow systems (e.g. Kepler, Taverna, Triana) and apply reasoning across these sets of records to infer new relationships between indirectly related data products:

• generate visualizations of the lineage of the data and its products
• dynamically infer customized views of provenance depending on the user’s requirements and privileges
• restrict access to specific data or processing steps
• Streamline the construction of publication or e-learning packages
Case Study

- Electrolytes are one of the primary fuel-cell components.

- The challenge is to determine the optimum combination of controllable parameters.

A logic view of the manufacture and testing process of Fuel-Cell Electrolyte.
Fig. 2. Provenance Model of the Electrolyte Manufacture and Analysis Process

SWRL

\[\text{IF (Experiment A includes Workflow B) AND (Workflow B contains Slip Batching C) AND (Slip Batching C hasInput Powder D) THEN (Experiment A hasInput Powder D)}\]
System Architecture

Provenance view

Publishing Interface

Provenance data

Two primary component of Shibboleth: IdP and SP

Fig. 3. System Architecture

Fig. 4. Authentication and Authorization System Architecture
Demonstration

Pink arrow means additional information could be generated

Fig. 5. A standard basic view

Fig. 6. Example policies and requests

| Fuel-Cell Researchers (CMM = "researcher") | Subject CMM="researcher" |
| Read All Views = "Permit" | Resource Resource="http://www.owl-ontologies.com/EM_ScientificProcess.owl/#SBViews" |
| | Action Action-id = "read" |
Demonstration (cont.)

Dark green arrows indicate links that can be collapsed manually back.

The complete provenance metadata for individual node on the upper panel will be displayed in the bottom panel.

Fig. 7. An expanded complete provenance view for the Researcher/Project Leader
Demonstration (inference)

The inferencing rule states that any products generated by one of the activities in the sequence is an output of the experiment.

Fig. 8. Demonstration of Provenance Inferencing
Discussion

• User feedback

• Limitation and Future work
  – Formal ontologies, such as DOLCE and BWW could technically enhance the reliability.
  – Extend the underlying model and the inferencing rules to support the data processing activities in the digital domain.
  – Determine a more streamlined mechanism for defining access policies and associating them with provenance relationships.
  – Support expansion down to multiple levels of detail.
  – The packages of components could be described, stored to institutional repositories and searched and retrieved for reuse.
Final Remarks
Conclusion

• These works clearly demonstrate integration Semantic Web and Provenance.

• Specifically:
  – Querying of workflow data through Semantic Web technologies
  – Applications of semantic inference applied toward provenance data
Future Work

• Provenance-Based Search on the web
• Integration of Provenance from heterogeneous sources
• Improved visualization interfaces
• Provenance-based information analysis
  – Could be used to infer properties of information, and serve as a basis for trust models