SimP: Secure Interoperable Multi-Granular Provenance Framework

Amani Abu Jabal ¹  Elisa Bertino ²

Purdue University, West Lafayette, USA
¹ aabujaba@purdue.edu, ² bertino@purdue.edu
Data Provenance

- Data provenance, one kind of metadata, which refers to the derivation history of a data object starting from its original sources.
  - Data object refers to data in any format (e.g., files, database records, or workflow templates).

- Comprehensive provenance infrastructure:
  - Multi-granular provenance model
  - Provenance queries
  - Security
  - Interoperability services
Existing Provenance Frameworks

- Provenance models tailored to specific applications:
  - Workflow-based provenance systems: Chimera [SSDBM’02], myGrid [ICSNW’04], and Karma [CCPE’08].
  - Process-based provenance systems: PreServ [AAAI’13]
  - OS-based provenance system: PASS [USENIX’06], and ES3 [IPAW’08].

- Standard Provenance Models (OPM and PROV).
  + Interoperable and Generic.
  - Not able to represent metadata about access control policies

- Ni’s model [SDM’09] focuses on access control policies.
  - It is not able to support different granularity levels

- The framework by Sultana and Bertino [JDM’15] is an initial comprehensive provenance infrastructure
  - Lacks interoperability services.
  - Not implemented nor integrated with an actual system.
Simp Framework

- Our provenance framework is composed of several components:
Provenance Model

- Main Entities in our model:
  - **Data**: data object (e.g. files)
  - **Processes**: activities which manipulate data
  - **Operations**: finer level of processes
  - **Actors**: actuator of data/processes (e.g. human)
  - **Environments**: system context parameters
  - **Access Controls**: policies placed at the time of data manipulation

- Our framework supports the specification of the provenance model in two representations: *relational* and *graph*. 
Beside the fundamental tables, there are:

- Lineages
- Communications
- Process Input/Output Data
- Operation Input/Output Data
- Delegations
Our graph model consists of 6 nodes and 12 types of edges.
Our framework supports interoperability with two standard provenance models: OPM and PROV.

The mapping ontology from PROV to SimP

<table>
<thead>
<tr>
<th></th>
<th>PROV</th>
<th>SimP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nodes</strong></td>
<td>Agent</td>
<td>Actor</td>
</tr>
<tr>
<td></td>
<td>Entity</td>
<td>Data</td>
</tr>
<tr>
<td></td>
<td>Activity</td>
<td>Process, Operation, WasPartOf</td>
</tr>
<tr>
<td><strong>Edges</strong></td>
<td>Used</td>
<td>Used</td>
</tr>
<tr>
<td></td>
<td>WasGeneratedBy</td>
<td>WasGeneratedBy</td>
</tr>
<tr>
<td></td>
<td>WasDerivedFrom</td>
<td>WasDerivedFrom</td>
</tr>
<tr>
<td></td>
<td>WasAssociatedWith</td>
<td>WasExecutedBy</td>
</tr>
<tr>
<td></td>
<td>WasInformedBy</td>
<td>WasInformedBy</td>
</tr>
<tr>
<td></td>
<td>WasAttributedTo</td>
<td>WasAttributedTo</td>
</tr>
<tr>
<td></td>
<td>ActedOnBehalfOf</td>
<td>ActedOnBehalfOf</td>
</tr>
</tbody>
</table>
Security & Granularity

- **Security:**
  - Access control policies
  - Restrict access to provenance storage

- **Granularity:**
  - Multi-granular Model
  - Granularity policies
Framework Implementation

- Provenance Storage:
  - Two types of storage: relational database (MySQL) and graph database (Neo4J).
  - Abstract storage interface: communicates with either MySQL adapter or Neo4J adaptor.

- Interoperability:
  - A service for converting from OPM or PROV (XML format) to SimP model.
Integration with a Scientific Data Management System

- Integrated with *Computational Research Infrastructure for Science (CRIS)*.
  - Used by a community of researchers at Purdue University

- For integration with CRIS:
  - Instrumenting component:
    - Use AOP to generate provenance logs (xml format)
  - Provenance Supplier:
    - Read provenance logs periodically
    - Convert into SimP XML
Conclusion and Future Work

- **SimP** – a comprehensive provenance framework
  - Includes a provenance model provided with relational and graph specifications
  - Interoperable with OPM and PROV
  - Supports multi-granular provenance
  - Supports security

- SimP is integrated with the scientific data management system “CRIS”.

- Future work:
  - Design and implement specialized query language for our framework
  - Investigate efficient compression techniques for our provenance model.
Thank you