Ontology-Driven Provenance Management in eScience: An Application in Parasite Research

Satya S. Sahoo  
*Wright State University - Main Campus*

D. Brent Weatherly

Raghava Mutharaju  
*Wright State University - Main Campus*

Pramod Anantharam  
*Wright State University - Main Campus*

Amit P. Sheth  
*Wright State University - Main Campus*, amit@sc.edu

See next page for additional authors

Follow this and additional works at: [https://corescholar.libraries.wright.edu/knoesis](https://corescholar.libraries.wright.edu/knoesis)

Part of the Bioinformatics Commons, Communication Technology and New Media Commons, Databases and Information Systems Commons, OS and Networks Commons, and the Science and Technology Studies Commons

Repository Citation

[https://corescholar.libraries.wright.edu/knoesis/669](https://corescholar.libraries.wright.edu/knoesis/669)

This Conference Proceeding is brought to you for free and open access by the The Ohio Center of Excellence in Knowledge-Enabled Computing (Kno.e.sis) at CORE Scholar. It has been accepted for inclusion in Kno.e.sis Publications by an authorized administrator of CORE Scholar. For more information, please contact library-corescholar@wright.edu.
Ontology-driven Provenance Management in eScience: An Application in Parasite Research

Satya S. Sahoo¹, D. Brent Weatherly², Raghava Mutharaju¹, Pramod Anantharam¹, Amit Sheth¹, Rick L. Tarleton²

¹Kno.e.sis Center, Wright State University;
²Center for Tropical and Emerging Diseases, University of Georgia

ODBASE2009
Vilamoura, Algarve-Portugal
November 05, 2009
Provenance in Parasite Research

Gene Knockout and Strain Creation*

Related Queries from Biologists

- Q2: List all groups in the lab that used a Target Region Plasmid?
- Q3: Which researcher created a new strain of the parasite (with ID = 66)?
- An experiment was not successful – has this experiment been conducted earlier? What were the results?

*Trypanosoma cruzi Semantic Problem Solving Environment Project, Courtesy of D.B. Weatherly and Flora Logan, Tarleton Lab, University of Georgia
• Provenance from the French word “provenir” describes the lineage or history of a data entity
• For Verification and Validation of Data Integrity, Process Quality, and Trust
• Issues in Provenance Management
  ➢ Provenance Modeling
  ➢ A Dedicated Query Infrastructure
  ➢ Practical Provenance Management Systems
Outline

• Provenance Modeling: Provenir → Parasite Experiment ontology

• Provenance Query Infrastructure

• Provenance Query Engine

• Evaluation Results

• Query Optimization: Materialized Provenance Views
Ontologies for Provenance Modeling

• Advantages of using Ontologies
  ➢ Formal Description: Machine Readability, Consistent Interpretation
  ➢ Use Reasoning: Knowledge Discovery over Large Datasets

• Problem: A gigantic, monolithic Provenance Ontology! – not feasible

• Solution: Modular Approach using a Foundational Ontology
Provenir Ontology

AGENT

DATA

PROCESS

has_agent

participates_in

Gene Name

Sequence Extraction

3' & 5' Region

Drug Resistant Plasmid

Plasmid Construction

Knockout Construct Plasmid

T. Cruzi sample

Transfection

Transfection Machine

Transfected Sample

Drug Selection

Selected Sample

Cell Cloning

Cloned Sample
Domain-specific Provenance: Parasite Experiment ontology

*Parasite Experiment ontology available at: http://wiki.knoesis.org/index.php/Trykipedia*
Outline

• Provenance Modeling: Provenir → Parasite Experiment ontology

• Provenance Query Infrastructure

• Provenance Query Engine

• Evaluation Results

• Query Optimization: Materialized Provenance Views
Classified Provenance Queries into Three Categories

- **Type 1: Querying for Provenance Metadata**
  - Example: *Which gene was used create the cloned sample with ID = 66?*

- **Type 2: Querying for Specific Data Set**
  - Example: *Find all knockout construct plasmids created by researcher Michelle using “Hygromycin” drug resistant plasmid between April 25, 2008 and August 15, 2008*

- **Type 3: Operations on Provenance Metadata**
  - Example: *Were the two cloned samples 65 and 46 prepared under similar conditions – compare the associated provenance information*
Four Query Operators – based on Query Classification

- **provenance ()** – Closure operation, returns the complete set of provenance metadata for input data entity
- **provenance_context()** - Given set of constraints defined on provenance, retrieves datasets that satisfy constraints
- **provenance_compare ()** - adapt the RDF graph equivalence definition
- **provenance_merge ()** - Two sets of provenance information are combined using the RDF graph merge
Answering Provenance Queries using \textit{provenance} \circ Operator
Outline

- Provenance Modeling: Provenir → Parasite Experiment ontology
- Provenance Query Infrastructure
- Provenance Query Engine
- Evaluation Results
- Query Optimization: Materialized Provenance Views
Provenance Query Engine

• Available as API for integration with provenance management systems
• Layer on top of a RDF Data Store (Oracle 10g), requires support for:
  o Rule-based reasoning
  o SPARQL query execution
• Input:
  o Type of provenance query operator: `provenance()`
  o Input value to query operator: `cloned sample 66`
  o User details to connect to underlying RDF store
Outline

• Provenance Modeling: Provenir →Parasite Experiment ontology

• Provenance Query Infrastructure

• Provenance Query Engine

• Evaluation Results

• Query Optimization: Materialized Provenance Views
Evaluation Results

- Queries expressed in SPARQL
- Datasets using real experiment data

<table>
<thead>
<tr>
<th>Query ID</th>
<th>Number of Variables</th>
<th>Total Number of Triples</th>
<th>Nesting Levels using OPTIONAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query 1: Target plasmid</td>
<td>25</td>
<td>84</td>
<td>4</td>
</tr>
<tr>
<td>Query 2: Plasmid_66</td>
<td>38</td>
<td>110</td>
<td>5</td>
</tr>
<tr>
<td>Query 3: Transfection attempts</td>
<td>67</td>
<td>190</td>
<td>7</td>
</tr>
<tr>
<td>Query 4: cloned_sample 66</td>
<td>67</td>
<td>190</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset ID</th>
<th>Number of RDF Inferred Triples</th>
<th>Total Number of RDF Triples</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS 1</td>
<td>2,673</td>
<td>3,553</td>
</tr>
<tr>
<td>DS 2</td>
<td>3,470</td>
<td>4,490</td>
</tr>
<tr>
<td>DS 3</td>
<td>4,988</td>
<td>6,288</td>
</tr>
<tr>
<td>DS 4</td>
<td>47,133</td>
<td>60,912</td>
</tr>
</tbody>
</table>
Evaluation Results

(a) Dataset Identifiers

(b) Query Identifiers
Outline

• Provenance Modeling: Provenir →Parasite Experiment ontology

• Provenance Query Infrastructure

• Provenance Query Engine

• Evaluation Results

• Query Optimization: Materialized Provenance Views
Query Optimization: Materialized Provenance Views

- Materializes a single logical unit of provenance
- Does not require query-rewriting
- View updates: addressed by characteristics of provenance
- Created using a memoization approach
Provenance Query Engine Architecture

**Query Optimizer**

**Transitive Closure**

**Provenance Query Engine**

**Input**
- Provenance Query Type
- Input Value

**Provenance Query Classification**
- Type of provenance query operator

**SPARQL Query Composer**
- PREFIX ro: <http://obofoundry.org/>
- CONSTRUCT
  
  ^p1 ro:has_participant trident:Ch...
  ^p1 ro:type provenir:process.

  WHERE
  { ?p1 ro:has_participant trident:Ch ;
    ro:type provenir:process . ...}.

**Data Source for Query Execution?**
- To underlying database

**Data Value Index**
- Index data value
- from new materialized graph

**Materialized Provenance Graphs**
- Data value found/not found

**Materialize Result Provenance Graph?**
- Yes
- No

**Oracle 10g RDF Database**

**Transitive Closure**
- ("process", "preceded_by")
Evaluation Results using Materialized Provenance Views

(a) Dataset Identifiers

(b) Query Identifiers
Provenance Management System for Parasite Research
Acknowledgement

• Flora Logan – The Wellcome Trust Sanger Institute, Cambridge, UK
• Priti Parikh – Kno.e.sis Center, Wright State University
• Roger Barga – Microsoft Research, Redmond
• Jonathan Goldstein – Microsoft Research, Redmond
Contact email: satyasahoo@gmail.com
Web: 
http://knoesis.wright.edu/researchers/satya/