Provenir ontology: Towards a Framework for eScience Provenance Management

Satya S. Sahoo, Amit P. Sheth
Kno.e.sis Center,
Wright State University

Microsoft eScience Workshop 2009
Pittsburgh, Oct 16
• Provenance: A Tale of Two Use Cases

• Provenance Ontologies: A Modular Approach

• Provenir: A Foundational Model of Provenance

• Provenance Query Infrastructure

• Application to Parasite Research
Cell Culture

- Extract

Glycoprotein Fraction

- Proteolysis

Glycopeptides Fraction

- Separation technique I

Glycopeptides Fraction

- PNGase

Peptide Fraction

- Separation technique II

Peptide Fraction

- Mass spectrometry

ms data

- Data reduction

ms peaklist

ms/ms data

- Data reduction

ms/ms peaklist

N-dimensional array

- Binning

Peptide list

Proteolytic enzyme
Gene Knockout and Strain Creation*

- 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
  - Interoperability
  - Consistent Modeling
  - Reduce Terminological Heterogeneity
• Provenance: A Tale of Two Use Cases

• Provenance Ontologies: A Modular Approach

• Provenir: A Foundational Model of Provenance

• Provenance Query Infrastructure

• Application to Parasite Research
• 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
• Provenance: A Tale of Two Use Cases

• Provenance Ontologies: A Modular Approach

• Provenir: A Foundational Model of Provenance

• Provenance Query Infrastructure

• Application to Parasite Research
Outline

Provenance: A Tale of Two Use Cases

Provenance Ontologies: A Modular Approach

Provenir: A Foundational Model of Provenance
Provenance Query Classification

Classified Provenance Queries into Three Categories

Type 1: Querying for Provenance Metadata

- Example: Which gene was used to create the cloned sample with ID = 65?

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
Provenance Query Operators

Our 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 Query Engine Architecture

Available as API for integration with provenance management systems. Input:

- Type of provenance query operator: `provenance()`
- Input value to
Provenance: A Tale of Two Use Cases

Provenance Ontologies: A Modular Approach

Provenir: A Foundational Model of Provenance
cruzi SPSE Provenance Management System
Provenir ontology as a foundational model for provenance
Extensible to model domain-specific provenance
- Parasite Experiment ontology
- Trident ontology
- ProPreO ontology

Query Infrastructure to support provenance modeled using Provenir ontology
Roger Barga – Microsoft Research, eScience
D. Brent Weatherly – Center for Tropical and Emerging Diseases, University of Georgia
Flora Logan – The Wellcome Trust Sanger Institute, Cambridge, UK
Raghava Mutharaju – Kno.e.sis Center, Wright State University
Pramod Anantharam – Kno.e.sis Center, Wright State University
Provenir ontology:
http://wiki.knoesis.org/index.php/Provenir_Ontology

Provenance Management in Parasite Research:
http://knoesis.wright.edu/library/resource.php?id=00712

Provenance Management Framework:
http://knoesis.wright.edu/research/semsci/application_domain/sem_prov/