Representing Microarray Experiment Metadata Using Provenance Models

Jim McCusker and Deborah L. McGuinness
Tetherless World Constellation, Rensselaer Polytechnic Institute
http://tw.rpi.edu
Observation

• experimental metadata (n):
  The information about an experiment that describes it in sufficient detail for someone to understand and reproduce the experiment.

• provenance in science (n):
  The information about an experiment that describes it in sufficient detail for someone to understand and reproduce the experiment.
Experimental metadata is provenance.
Examples: Provenance Representations

All of these are representations of provenance.
Examples: Provenance Representations

• Computational Workflows

All of these are representations of provenance.
Examples: Provenance Representations

- Computational Workflows
- Experimental Workflows

All of these are representations of provenance.
Examples: Provenance Representations

- Computational Workflows
- Experimental Workflows
- Biospecimen Management

All of these are representations of provenance.
Examples: Provenance Representations

• Computational Workflows
• Experimental Workflows
• Biospecimen Management
• Data Sources, Ownership & Attribution

All of these are representations of provenance.
Hypothesis

• If experimental metadata is provenance:
  • Then a good model of provenance can be used to encode experimental metadata as well as a domain-specific metadata.
  • Then a general model of provenance can model the entirety of experimental metadata.
• We try to evaluate this by mapping microarray metadata onto provenance graphs.
Why a generic provenance model?
An Integrated View of Provenance
An Integrated View of Provenance

Transformation

Normalization

Scan

Hybridization

Labeled Extract → Extract → Cell Line
An Integrated View of Provenance

Transformation

Normalization

Scan

Hybridization

Labeled Extract

Extract

Cell Line

Section

Biopsy

Participant
An Integrated View of Provenance

Transformation

Normalization

Scan

Hybridization

Labeled Extract

Extract

Cell Line

Section

Biopsy

Participant

Clinical Diagnosis
An Integrated View of Provenance
An Integrated View of Provenance

Transformation
Normalization
Scan
Hybridization
Labeled Extract
Extract
Cell Line
Section
Biopsy
Participant
Clinical Diagnosis
DOB
Histology Image
Slide
An Integrated View of Provenance

Transformation
Normalization
Scan
Hybridization
Labeled Extract
Extract
Cell Line
Section
Biopsy
Participant
Clinical Diagnosis
DOB
Histology Image
Slide
An Integrated View of Provenance
An Integrated View of Provenance

Transformation → Normalization → Scan → Hybridization → Labeled Extract → Extract → Cell Line → Section → Biopsy → Participant

- Lab Results
- Histology Image
- Blood Sample
- Clinical Diagnosis
- DOB
- Adverse Events
Related Work, ABC Soup

- MAGE (MicroArray and Gene Expression Object Model)
- MAGE-TAB: Tab-delimited spreadsheet of MAGE.
  - IDF (Investigation Design Format)
  - SDRF (Sample and Data Relationship Format)
- MGED (Microarray Gene Expression Data) Ontology and MGED Society
- OPM (Open Provenance Model)
- PML (Proof Markup Language)
General Models of Provenance: Open Provenance Model
General Models of Provenance: Proof Markup Language
Methods Overview

- MAGE-TAB IDF
- MAGE-TAB SDRF
Methods Overview

- MAGE-TAB IDF
- MAGE-TAB SDRF
- MAGE-TAB to MAGE-RDF
- MAGE-RDF

Relationships:
- Used(IDF) from MAGE-TAB to MAGE-TAB SDRF
- Used(SDFR) from MAGE-TAB SDRF to MAGE-TAB to MAGE-RDF
- WasGeneratedBy(RDF) from MAGE-TAB to MAGE-RDF
Methods Overview

MAGE-TAB IDF

MAGE-TAB to MAGE-RDF

MAGE-TAB SDRF

mageprov.rules

Jena Rule Engine

mageprovenance.owl

Used(IDF)

Used(SDRF)

Used(RDF)

WasGeneratedBy(RDF)

Used(Rules)

Used(OWL)

Used(RDF)
Methods Overview

MAGE-TAB

IDF

Used(IDF)

MAGE-TAB

SDRF

Used(SDRF)

MAGE-TAB

to

MAGE-RDF

WasGeneratedBy(RDF)

MAGE-RDF

PML & OPM

WasGeneratedBy(RDF)

mageprov.rules

Used(Rules)

mageprovenance.owl

Used(OWL)

Jena Rule Engine

Used(RDF)
MAGE-TAB 2 MAGE-RDF: an RDF Version of MAGE

• Used Limpopo to parse MAGE-TAB
• Big thanks to HCLS SIG for input
• (Small) issues with MGED Ontology:
  • Missing ProtocolApplication, some key properties.
  • Cannot perform inferencing without ignoring owl:someValuesFrom.
    • Contains 234 owl:Classes with 193 owl:someValuesFrom restrictions.
    • Should be using integrity constraint instead.
• http://magetab2rdf.googlecode.com
Methods: MAGE-RDF to Open Provenance Model

- 16 rules in Jena Inference Language
- 15 subClass mappings
- 10 individuals (Roles)
- 2 subProperty mappings

Example Rule:

```prolog
[opm.generated.sample:
    (?exp rdf:type mged:Experiment),
    (?exp mage:has_protocol_application ?pa),
    makeTemp(?gen),
    (?pa mged:has_protocol ?protocol),
    (?pa mage:has_derivative ?sample)
  ->
    (?gen rdf:type opm:WasGeneratedBy),
    (?gen opm:cause ?protocol),
    (?gen opm:account ?exp),
    (?gen opm:effect ?sample),
    (?sample rdf:type opm:Artifact)]
```
Methods: MAGE-RDF to Proof Markup Language

- 4 rules in Jena Inference Language
- 14 subClass mappings
- 2 subProperty mappings

Example Rule:

```plaintext
[pml.protocolApplication:  
(?pa rdf:type mage:ProtocolApplication),  
(?pa mage:has_derivation_source ?derivation),  
(?pa mged:has_protocol ?protocol),  
(?pa mage:has_derivative ?derived),  
(?pa mged:has_protocol ?protocol),  
(?derivedNodeSet pmlj:hasConclusion ?derived),  
(?derivationNodeSet pmlj:hasConclusion ?derivation),  
makeTemp(?antecedentList)  
->  
(?pa rdf:type pmlj:InferenceStep),  
(?pa pmlj:hasInferenceRule ?protocol),  
(?pa pmlj:hasIndex 1),  
(?derivedNodeSet pmlj:isConsequentOf ?pa),  
(?pa pmlj:hasAntecedentList ?antecedentList),  
(?antecedentList rdf:type pmlj:NodeSetList),  
(?antecedentList ds:first ?derivationNodeSet)
]```
# Evaluation: Declarative Mapping

## Pros:
- Easy to understand.
- Easy to validate.
- Little programming involved.
- Many languages.
- Easy integration with existing ontologies.
- Powerful rule languages available.

## Cons:
- “Unclean” representations.
- More difficult to debug.
- Many languages.
- Logic programming not everyone's cup of tea.
Evaluation:
Expressiveness & Coverage
Evaluation: Expressiveness & Coverage

Missing in OPM:

• See below.
Evaluation:
Expressiveness & Coverage

Missing in OPM:
• See below.

Missing in PML:
• Parameters
• Parameter Values (could have been variable mappings)
Evaluation: Expressiveness & Coverage

Missing in OPM:
• See below.

Could have added to OPM and PML:
• Publications
• Time
• Comments

Missing in PML:
• Parameters
• Parameter Values (could have been variable mappings)
Evaluation: Expressiveness & Coverage

**Missing in OPM:**
- See below.

**Could have added to OPM and PML:**
- Publications
- Time
- Comments

**Missing in PML:**
- Parameters
- Parameter Values (could have been variable mappings)

MAGE comments are structured name-value pairs. These would have to be represented as PML Information and OPM Artifacts.
Evaluation:
PML Visualization
Evaluation: PML Visualization
Evaluation:
OPM Visualization
Evaluation: OPM Visualization
Future Work

• Integration with:
  • Tissue management tools
  • Clinical data sources
  • Workflow engines (Taverna already exports OPM)

• Use case-based evaluation of sufficiency.
Conclusions

• Experimental metadata is provenance.
• All experimental provenance can benefit from a common model.
• MAGE-provenance means 9,975 assays ready for conversion to a common model.
• OPM provides slightly better modeling coverage and workflow integration.
• PML provides better reasoning coverage and theorem prover integration.
Acknowledgements
and References

• Tetherless World (Deborah McGuinness et al.)
• Krauthammer Lab (Michael Krauthammer et al.)
• W3C Health Care and Life Sciences SIG.
• LIMPOPO: limpopo.sourceforge.net
• MAGE-TAB: magetab.sourceforge.net
• MGED Society: www.mged.org
• PML: www.inference-web.org
• OPM: www.openprovenance.org
• Jiao Tao, Li Ding, Deborah L. McGuinness, Instance Data Evaluation for Semantic Web-Based Knowledge Management Systems, 42th Hawaii International Conference on System Sciences (HICSS-42), January 2009.