Formal Verification of Data Provenance Records

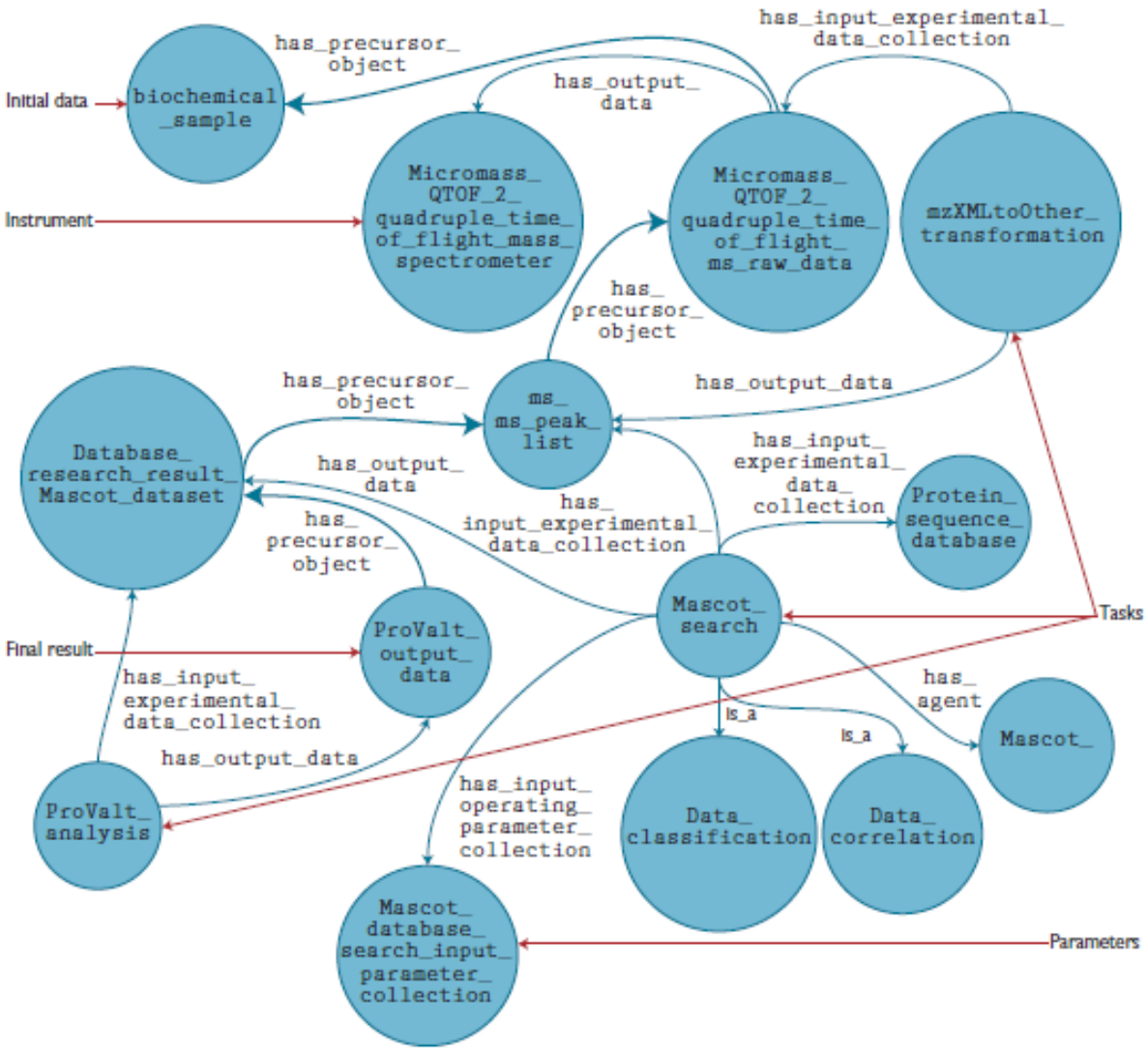
Szymon Klarman[1], Stefan Schlobach[1] and Luciano Serafini[2]

[1] Knowledge Representation & Reasoning Group
VU University Amsterdam

[2] Data & Knowledge Management Group
FBK Trento

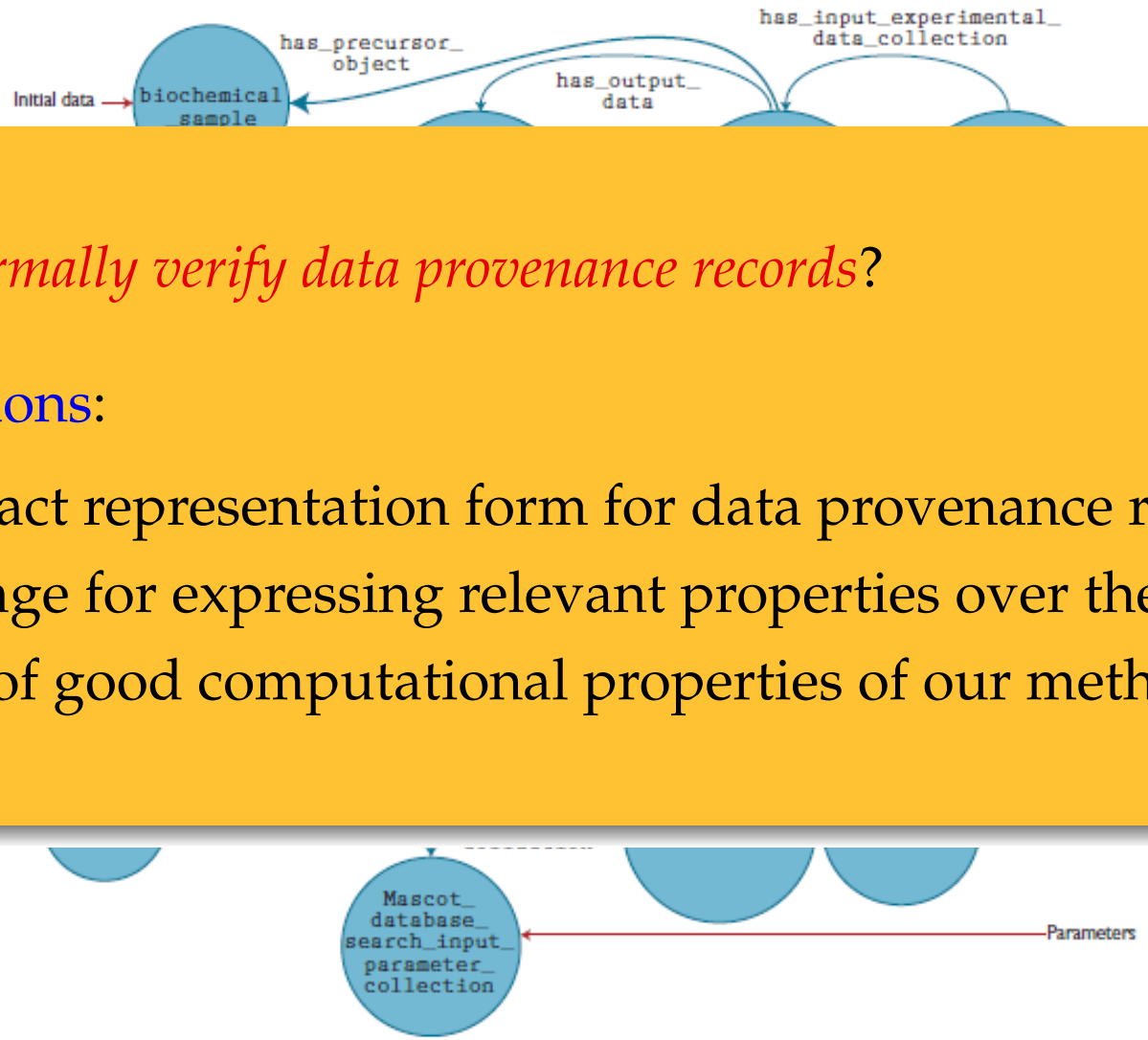
The 11th International Semantic Web Conference (ISWC-12)

Problem



Sahoo, S., Sheth, A., Henson, C. **Semantic Provenance for eScience: Managing the Deluge of Scientific Data.** In *IEEE Internet Computing 12(4)*, 2008.

Problem



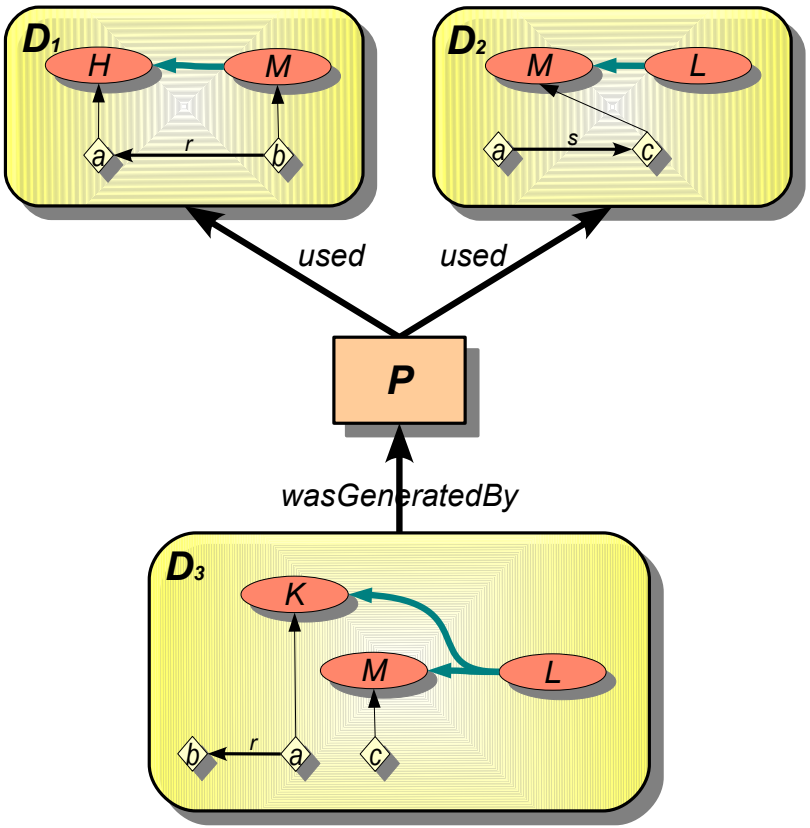
How to *formally verify data provenance records*?

Contributions:

- an abstract representation form for data provenance records,
- a language for expressing relevant properties over the records,
- a proof of good computational properties of our method.

Sahoo, S., Sheth, A., Henson, C. **Semantic Provenance for eScience: Managing the Deluge of Scientific Data.** In *IEEE Internet Computing 12(4)*, 2008.

Data provenance records



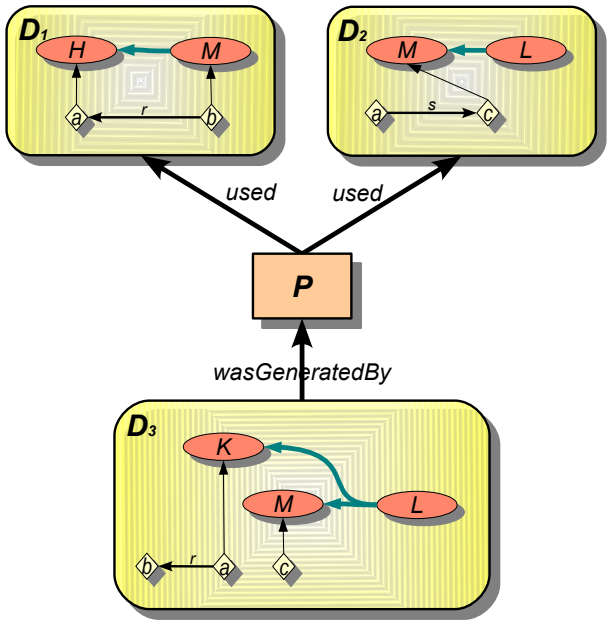
Provenance graphs:

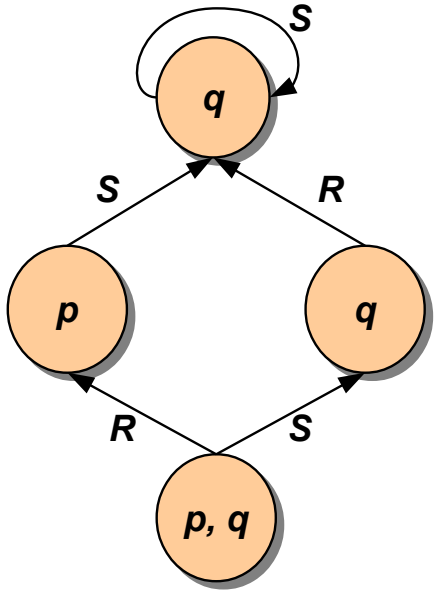
- *process* nodes: P
- *data artifact* nodes: D_1, D_2, D_3 (each corresponding to an OWL/RDF(S) dataset)
- edges labeled with *relation names*, e.g.: *wasGeneratedBy*, *used*.
- directed, acyclic, finite.

L. Moreau, et al. **The open provenance model – core specification.** In *Future Generation Computer Systems 27*, 2010.

Provenance Specification Logic

Provenance graphs are very similar to *finite-state transition systems*.





- natural to analyze using *Propositional Dynamic Logic* (PDL)
- *Provenance Specification Logic* = PDL with [propositions $\mapsto$ *queries*]
- basic reasoning task is *model-checking*

Validation

Modeling the test queries from *The First Provenance Challenge*, e.g.:

> *Q: Find all output averaged images of softmean (average) procedures, where the warped images taken as input were align warp'ed using a twelfth order nonlinear 1365 parameter model, i.e. where softmean was preceded in the workflow, directly or indirectly, by an align warp procedure with argument -m 12.*

Such *two-dimensional* queries can be conveniently captured in our logic.

L. Moreau, et al. **Special issue: The First Provenance Challenge.** In *Concurrency and Computation: Practice and Experience 20*, 2008.

Summary

Our *Provenance Specification Logic* is:

- generic: independent of particular data/provenance representation languages,
- declarative: allows for abstract formulation of useful properties over data provenance graphs,
- well-behaved: combines traditional model-checking techniques with standard Semantic Web reasoning.