Mapping the NRC Dataflow Model to the Open Provenance Model

Natalia Kwasnikowska
and
Jan Van den Bussche

Hasselt University and Transnational University of Limburg
Belgium

DILS
June 25-27, 2008

NRC Dataflow Model

- Nested Relational Calculus
  - complex-data flow
  - service names
- Formal representation of past executions
- Subvalue provenance in a past execution
- Different dataflows stored in a repository
  - sub-dataflows
  - late binding of service names to external services

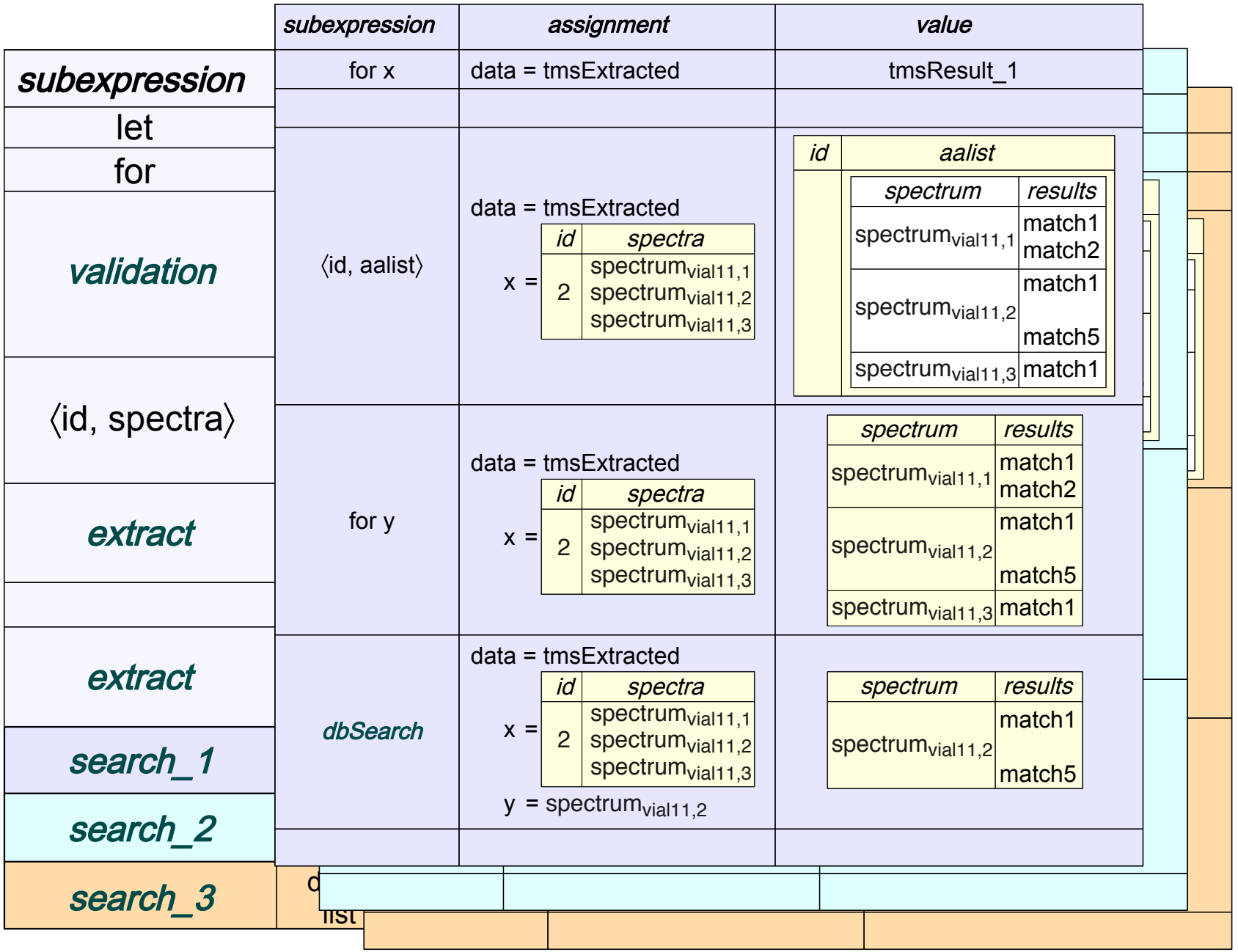
Open Provenance Model

- An OPM graph represents a past execution
- Artifact, process and agent nodes
  - 5 types of edges
- Account membership of nodes and edges
- Alternate accounts of a past execution

Mapping

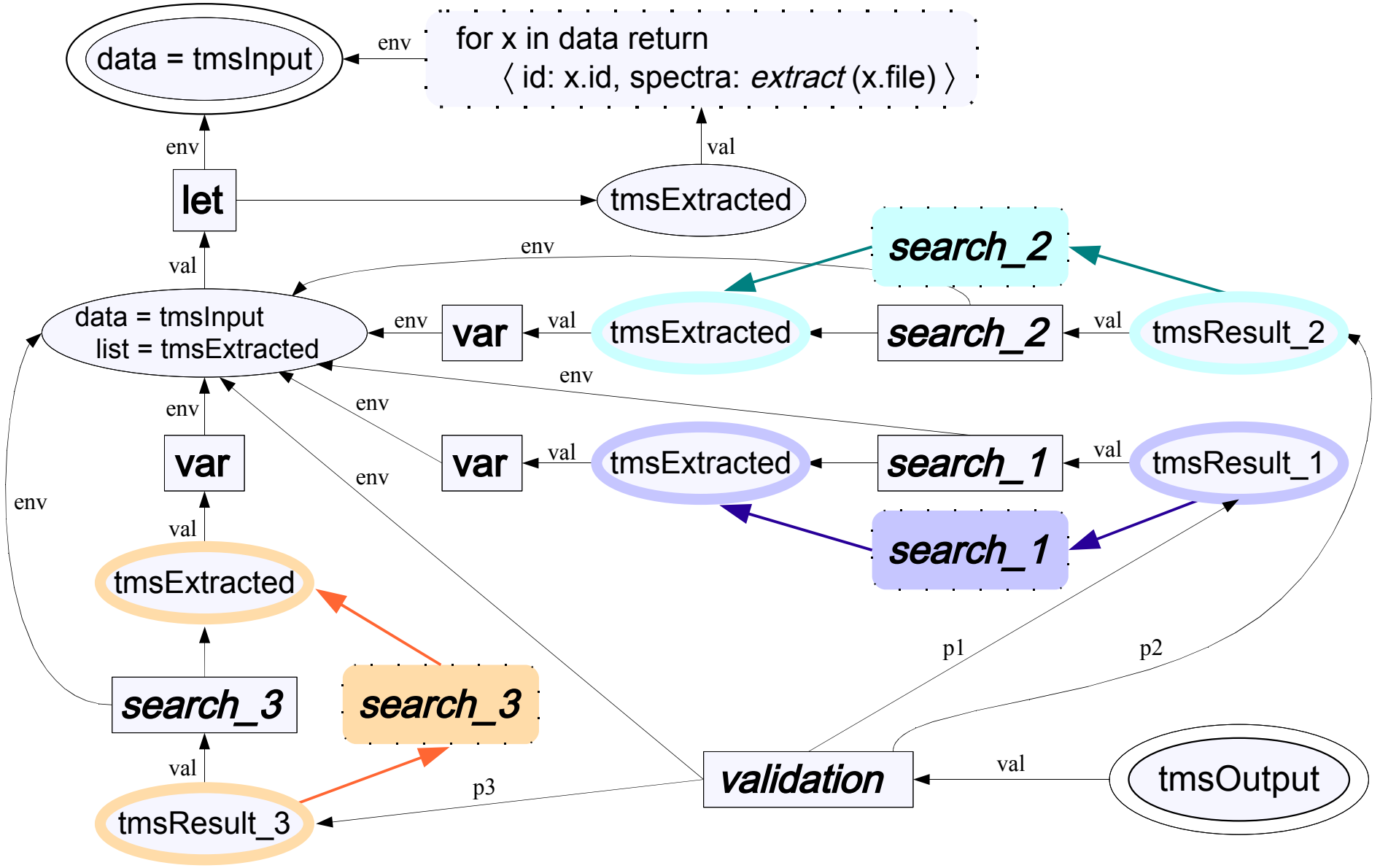
- Validation of both models
- How much of an NRC dataflow execution can be represented as an OPM graph?
- What about subvalue provenance?

Record of a Past Execution of *identify*(tmsInput)

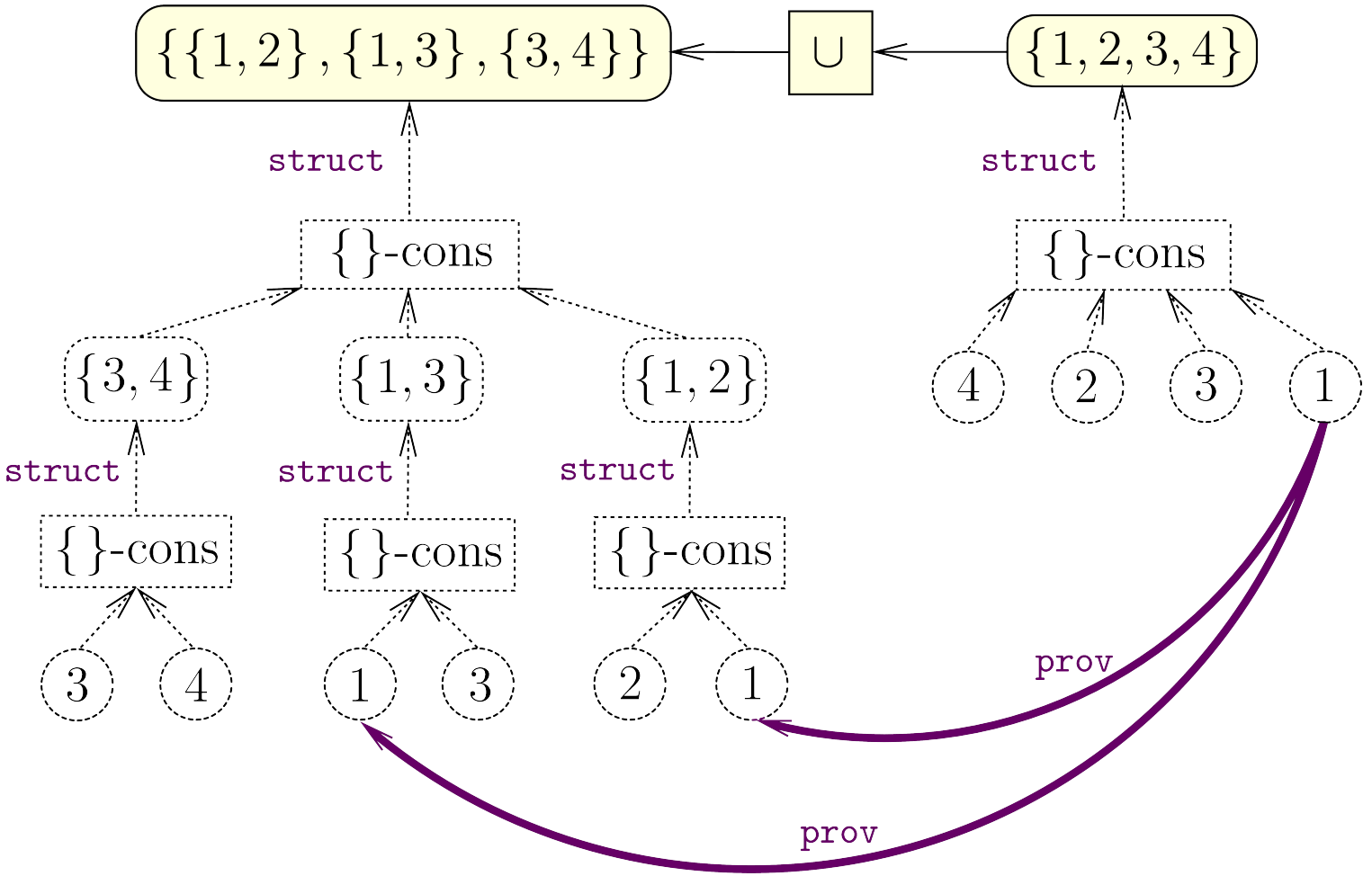


OPM graph

```
dataflow identify(data: TMSdata): ProteinCandidateList is
  let list := for x in data return
                ⟨ id: x.id, spectra: extract(x.file) ⟩
  in validation(search_1(list), search_2(list), search_3(list))
```



Subvalue Provenance



Mapping the NRC Dataflow Model to the Open Provenance Model

Natalia Kwasnikowska
and
Jan Van den Bussche

11

universiteit hasselt

transnationale UNIVERSITEIT LIMBURG