

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

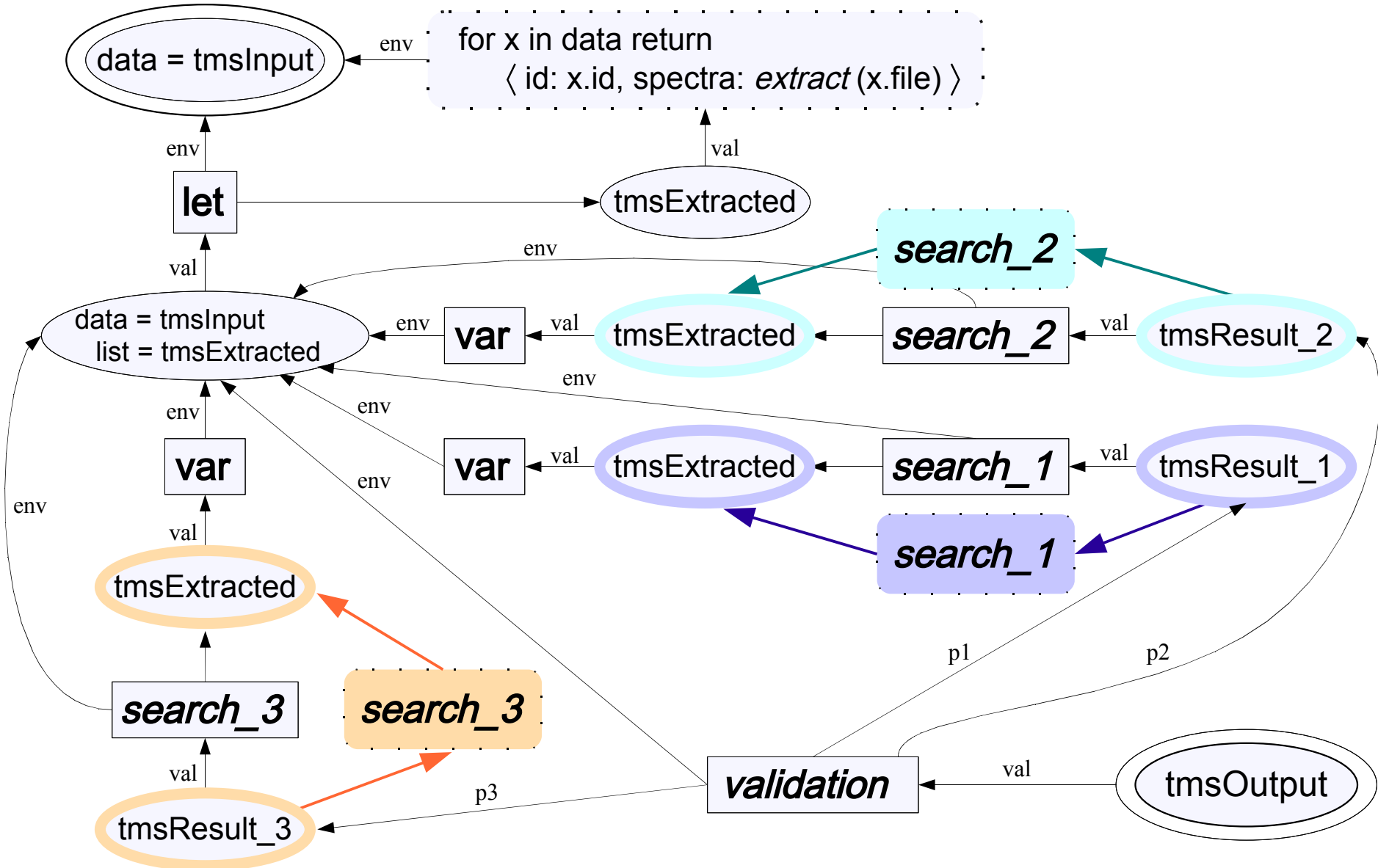
- Validation of both models
- How much of an NRC dataflow execution can be represented as an OPM graph?
- What about subvalue provenance?

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

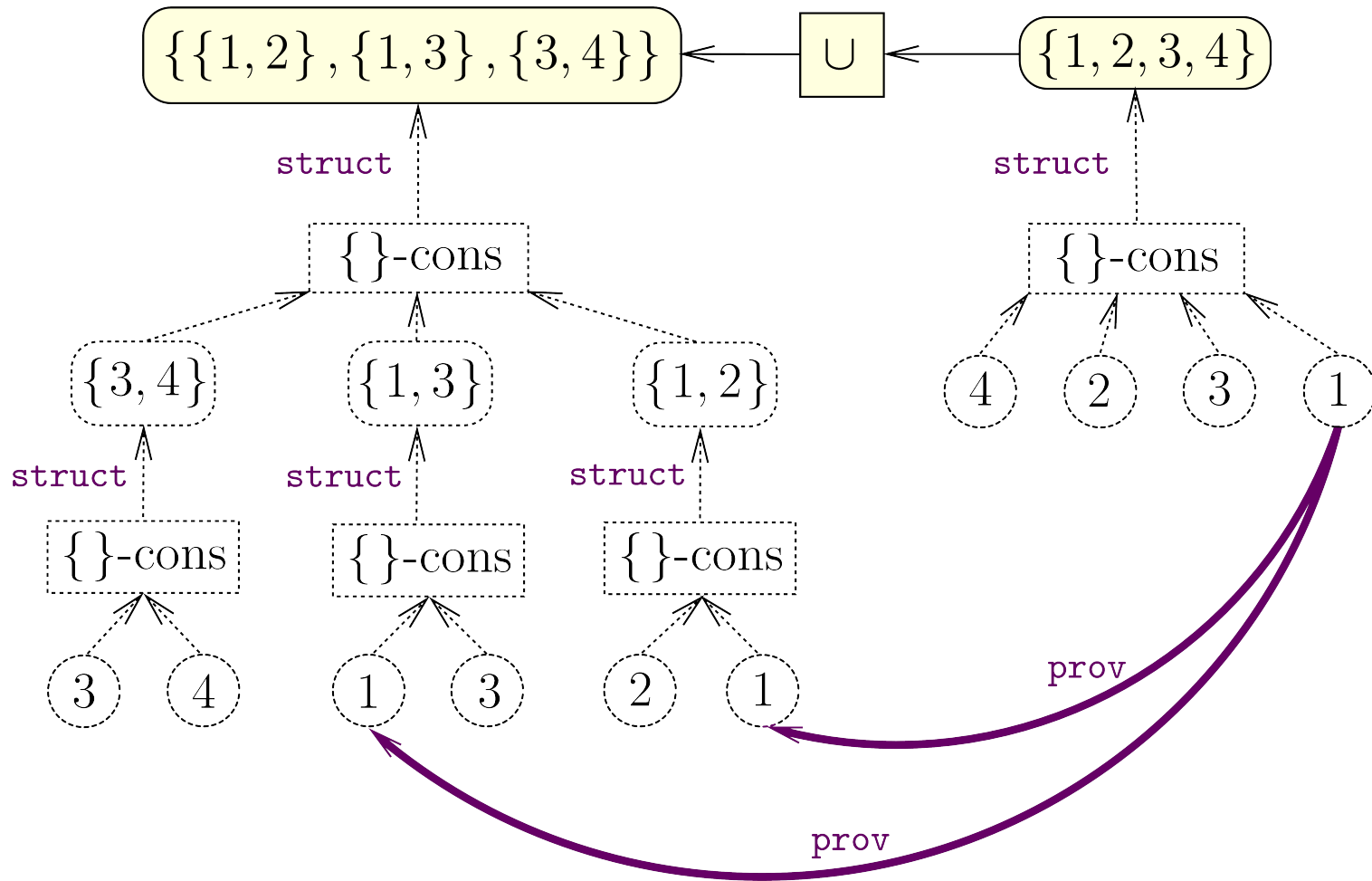
	<i>subexpression</i>	<i>assignment</i>	<i>value</i>																			
<i>subexpression</i>	for x	data = tmsExtracted	tmsResult_1																			
let																						
for																						
<i>validation</i>	⟨id, alist⟩	data = tmsExtracted x = <table border="1"> <thead> <tr> <th><i>id</i></th> <th><i>spectra</i></th> </tr> </thead> <tbody> <tr> <td>2</td> <td>spectrum_{vial11,1} spectrum_{vial11,2} spectrum_{vial11,3}</td> </tr> </tbody> </table>	<i>id</i>	<i>spectra</i>	2	spectrum _{vial11,1} spectrum _{vial11,2} spectrum _{vial11,3}	<table border="1"> <thead> <tr> <th><i>id</i></th> <th colspan="2"><i>alist</i></th> </tr> <tr> <td></td> <th><i>spectrum</i></th> <th><i>results</i></th> </tr> </thead> <tbody> <tr> <td></td> <td>spectrum_{vial11,1}</td> <td>match1 match2</td> </tr> <tr> <td></td> <td>spectrum_{vial11,2}</td> <td>match1 match5</td> </tr> <tr> <td></td> <td>spectrum_{vial11,3}</td> <td>match1</td> </tr> </tbody> </table>	<i>id</i>	<i>alist</i>			<i>spectrum</i>	<i>results</i>		spectrum _{vial11,1}	match1 match2		spectrum _{vial11,2}	match1 match5		spectrum _{vial11,3}	match1
<i>id</i>	<i>spectra</i>																					
2	spectrum _{vial11,1} spectrum _{vial11,2} spectrum _{vial11,3}																					
<i>id</i>	<i>alist</i>																					
	<i>spectrum</i>	<i>results</i>																				
	spectrum _{vial11,1}	match1 match2																				
	spectrum _{vial11,2}	match1 match5																				
	spectrum _{vial11,3}	match1																				
⟨id, spectra⟩																						
<i>extract</i>	for y	data = tmsExtracted x = <table border="1"> <thead> <tr> <th><i>id</i></th> <th><i>spectra</i></th> </tr> </thead> <tbody> <tr> <td>2</td> <td>spectrum_{vial11,1} spectrum_{vial11,2} spectrum_{vial11,3}</td> </tr> </tbody> </table>	<i>id</i>	<i>spectra</i>	2	spectrum _{vial11,1} spectrum _{vial11,2} spectrum _{vial11,3}	<table border="1"> <thead> <tr> <th><i>spectrum</i></th> <th><i>results</i></th> </tr> </thead> <tbody> <tr> <td>spectrum_{vial11,1}</td> <td>match1 match2</td> </tr> <tr> <td>spectrum_{vial11,2}</td> <td>match1 match5</td> </tr> <tr> <td>spectrum_{vial11,3}</td> <td>match1</td> </tr> </tbody> </table>	<i>spectrum</i>	<i>results</i>	spectrum _{vial11,1}	match1 match2	spectrum _{vial11,2}	match1 match5	spectrum _{vial11,3}	match1							
<i>id</i>	<i>spectra</i>																					
2	spectrum _{vial11,1} spectrum _{vial11,2} spectrum _{vial11,3}																					
<i>spectrum</i>	<i>results</i>																					
spectrum _{vial11,1}	match1 match2																					
spectrum _{vial11,2}	match1 match5																					
spectrum _{vial11,3}	match1																					
<i>extract</i>																						
<i>search_1</i>	<i>dbSearch</i>	data = tmsExtracted x = <table border="1"> <thead> <tr> <th><i>id</i></th> <th><i>spectra</i></th> </tr> </thead> <tbody> <tr> <td>2</td> <td>spectrum_{vial11,1} spectrum_{vial11,2} spectrum_{vial11,3}</td> </tr> </tbody> </table>	<i>id</i>	<i>spectra</i>	2	spectrum _{vial11,1} spectrum _{vial11,2} spectrum _{vial11,3}	<table border="1"> <thead> <tr> <th><i>spectrum</i></th> <th><i>results</i></th> </tr> </thead> <tbody> <tr> <td>spectrum_{vial11,2}</td> <td>match1 match5</td> </tr> </tbody> </table>	<i>spectrum</i>	<i>results</i>	spectrum _{vial11,2}	match1 match5											
<i>id</i>	<i>spectra</i>																					
2	spectrum _{vial11,1} spectrum _{vial11,2} spectrum _{vial11,3}																					
<i>spectrum</i>	<i>results</i>																					
spectrum _{vial11,2}	match1 match5																					
<i>search_2</i>		y = spectrum _{vial11,2}																				
<i>search_3</i>																						

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