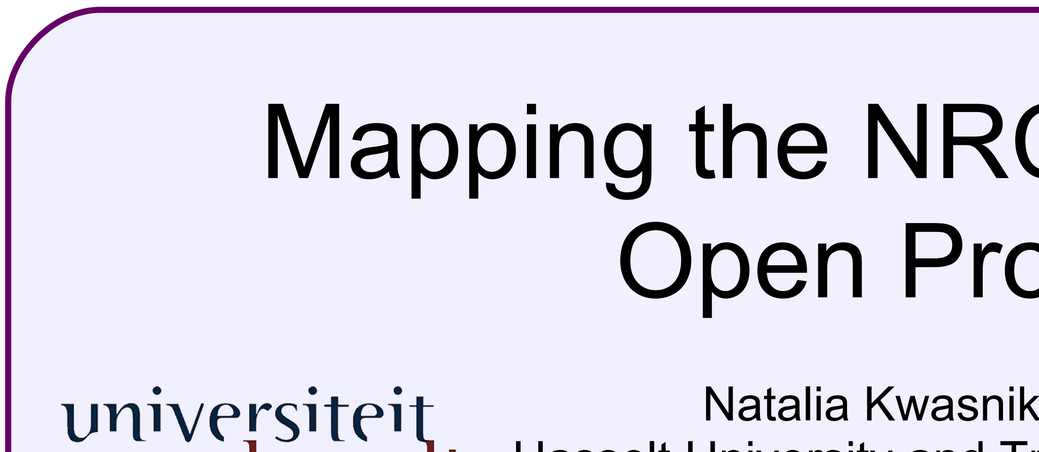
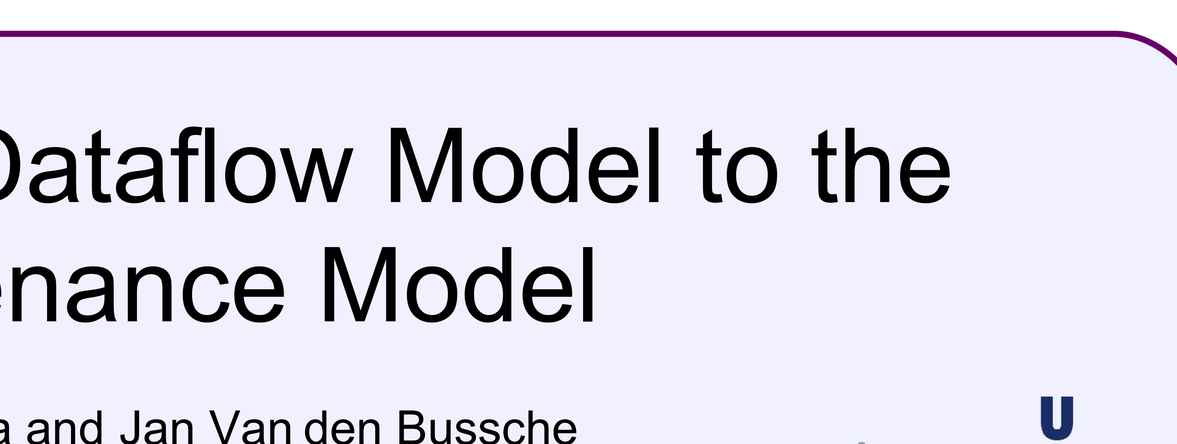
Mapping the NRC Dataflow Model to the Open Provenance Model

Natalia Kwasnikowska and Jan Van den Bussche
Hasselt University and Transnational University of Limburg, Belgium

Nested Relational Calculus Dataflow Model



Open Provenance Model (OPM)



Given a set of **raw data** produced in a proteomics experiment, generate a list of possibly identified **proteins**

{ ⟨ protID: **ProteinID** , prob: **Number** , evidence: { Peptide } ⟩ }

```
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))

dataflow search (data: TMSextracted): AAlist is
  for x in data return
    ⟨ id: x.id, aalist: for y in x.spectra return dbSearch(y) ⟩
```

validate — extract — 1 — 2 — 3 — mascot — sequest — xtandem

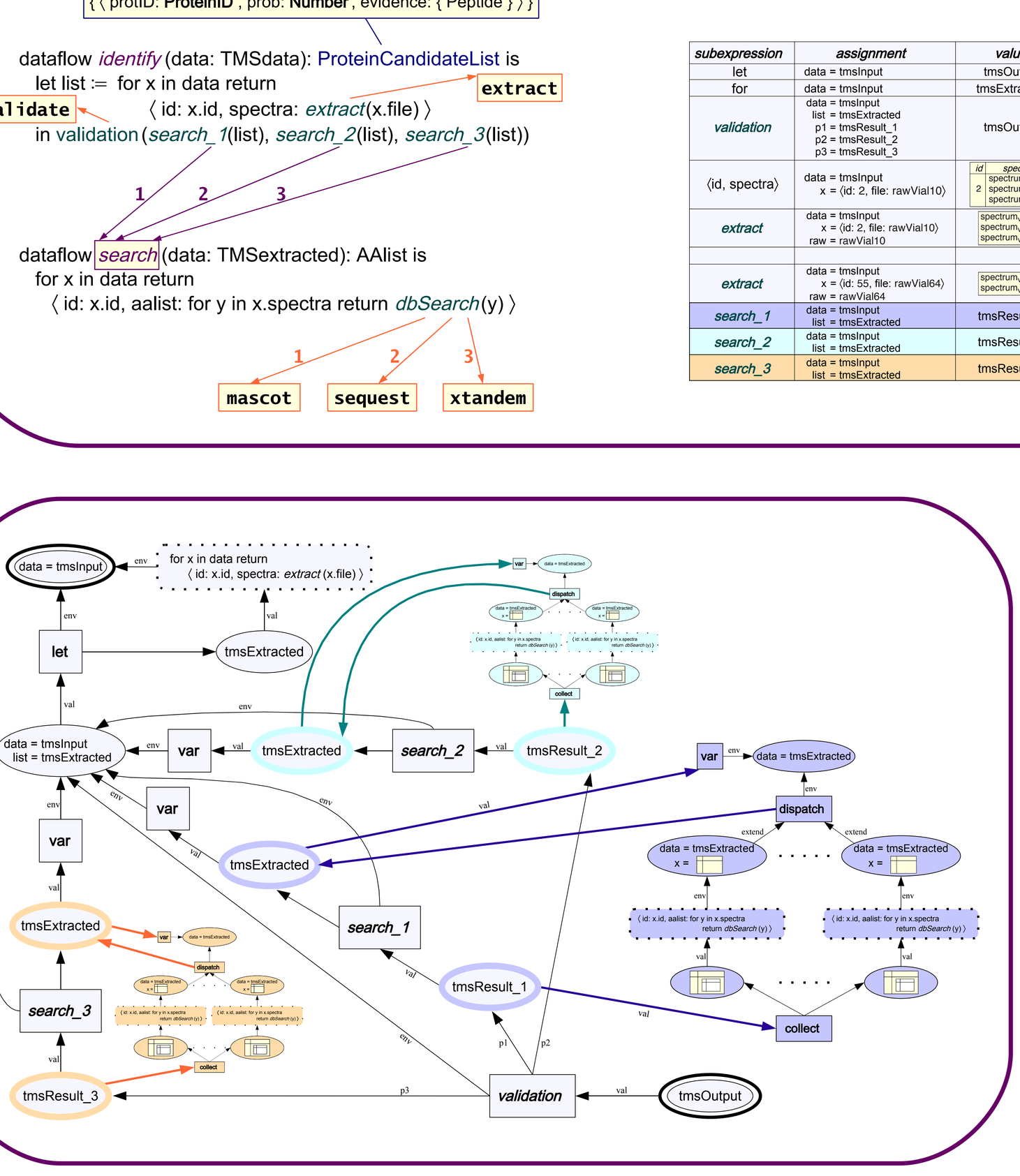
id	file
1	rawVial10
2	rawVial11
42	rawVial51
55	rawVial64

tmsInput → *identify* → tmsOutput

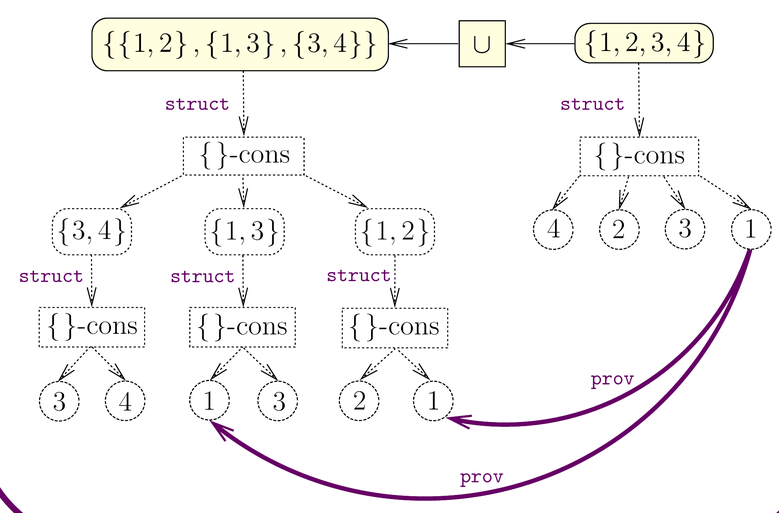
protID	prob	evidence		
		peptide	*score*	*spectrum*
protein$_1$	99	pep$_1$	9	spectrum$_{vial10,5}$
		pep$_2$	7	spectrum$_{vial23,2}$
protein$_2$	96	pep$_8$	9	spectrum$_{vial51,3}$
		pep$_1$	8	spectrum$_{vial10,5}$

subexpression	*assignment*	*value*
let	data = tmsInput	tmsOutput
for	data = tmsInput	tmsExtracted
validation	data = tmsInput list = tmsExtracted p1 = tmsResult_1 p2 = tmsResult_2 p3 = tmsResult_3	tmsOutput
⟨id, spectra⟩	data = tmsInput x = ⟨id: 2, file: rawVial10⟩	id: 2 → spectrum$_{vial10,1}$, spectrum$_{vial10,2}$, spectrum$_{vial10,3}$
extract	data = tmsInput x = ⟨id: 2, file: rawVial10⟩ raw = rawVial10	spectrum$_{vial10,1}$ spectrum$_{vial10,2}$ spectrum$_{vial10,3}$
extract	data = tmsInput x = ⟨id: 55, file: rawVial64⟩ raw = rawVial64	spectrum$_{vial64,1}$ spectrum$_{vial64,2}$
search_1	data = tmsInput list = tmsExtracted	tmsResult_1
search_2	data = tmsInput list = tmsExtracted	tmsResult_2
search_3	data = tmsInput list = tmsExtracted	tmsResult_3

subexpression	*assignment*	*value*
for x	data = tmsExtracted	tmsResult_1
⟨id, aalist⟩	data = tmsExtracted x = id: 2 → spectrum$_{vial11,1}$, spectrum$_{vial11,2}$, spectrum$_{vial11,3}$	id: 2 → aalist: spectrum$_{vial11,1}$ → match1, match2 spectrum$_{vial11,2}$ → match1, match5 spectrum$_{vial11,3}$ → match1
for y	data = tmsExtracted x = id: 2 → spectrum$_{vial11,1}$, spectrum$_{vial11,2}$, spectrum$_{vial11,3}$	spectrum$_{vial11,1}$ → match1, match2 spectrum$_{vial11,2}$ → match1, match5 spectrum$_{vial11,3}$ → match1
dbSearch	data = tmsExtracted x = id: 2 → spectrum$_{vial11,1}$, spectrum$_{vial11,2}$, spectrum$_{vial11,3}$ y = spectrum$_{vial11,2}$	spectrum$_{vial11,2}$ → match1, match5
	y = spectrum$_{vial42,51}$	
	y = spectrum$_{vial10,3}$	



Subvalue Provenance



References

Kwasnikowska, N., Van den Bussche, J.: Mapping the NRC Dataflow Model to the Open Provenance Model. Second International Provenance and Annotation Workshop (IPAW), June 17-18, Salt Lake City, UT

Moreau, L., Freire, J., Futrelle, J., McGrath, R., Myers, J., Paulson, P.: The open provenance model. Technical Report 14979, University of Southampton, School of Electronics and Computer Science (2007)

Hidders, J., Kwasnikowska, N., Sroka, J., Tyszkiewicz, J., Van den Bussche, J.: A formal model of dataflow repositories. In Cohen-Boulakia, S., Tannen, V., eds.: DILS. LNCS Vol. 4544, Springer (2007) 105–121

Buneman, P., Naqvi, S., Tannen, V., Wong, L.: Principles of programming with complex objects and collection types. Theoretical Computer Science 149 (1995) 3–48

Dumont, D., Noben, J., Raus, J., Stinissen, P., Robben, J.: Proteomic analysis of cerebrospinal fluid from multiple sclerosis patients. Proteomics 4(7) (2004) 2117– 2124