MANCHESTER 1824

The University of Manchester

iSPIDER

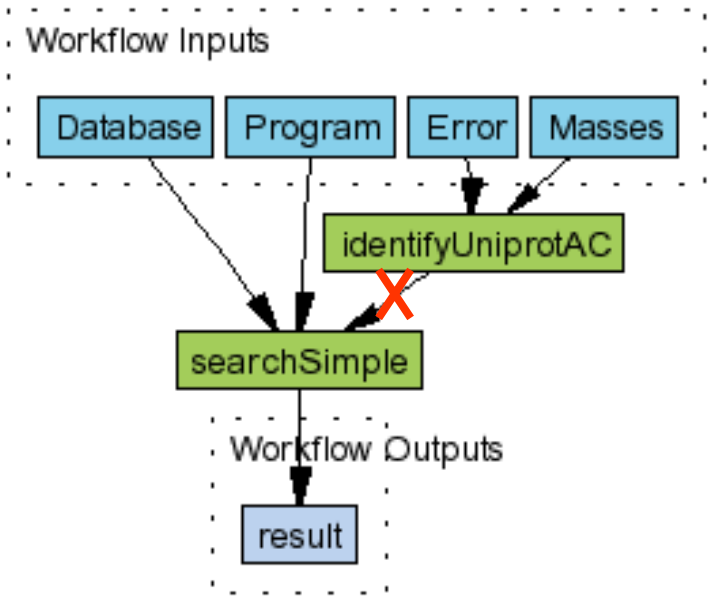
On characterising and identifying mismatches in scientific workflows

Khalid Belhajjame, Suzanne M. Embury, and Norman W. Paton
School of Computer Science
University of Manchester

Scientific workflow

- A scientific workflow is a series of analysis operations connected using data links
- Analysis operations are supplied by independently developed web services

→ Connected parameters can be mismatched



Objective: to characterise mismatches in scientific workflows and provide support for their automatic detection

Outline

- ✓ Scientific workflows
- Ontologies for describing operation parameters
- Classes of mismatches
- Evaluation

Ontologies

- **Domain ontology**: captures information about the application domains covered by operation parameters, e.g., Protein_record and DNA_sequence

$$domain: OP \times (IN \cup OUT) \rightarrow \theta_{domain}$$

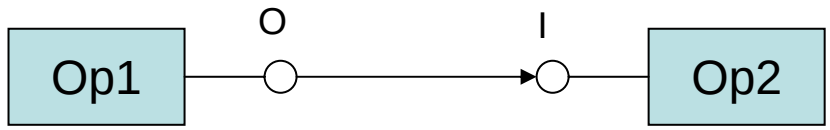
- **Representation ontology**: describes the format of data, e.g., Uniprot_record and Fasta_record

$$represent: OP \times (IN \cup OUT) \rightarrow \theta_{represent}$$

- **Extent ontology**: defines the scope of values of operation parameters, e.g., SwissProt_datastore

$$extent: OP \times (IN \cup OUT) \rightarrow \theta_{extent}$$

Classes of mismatches



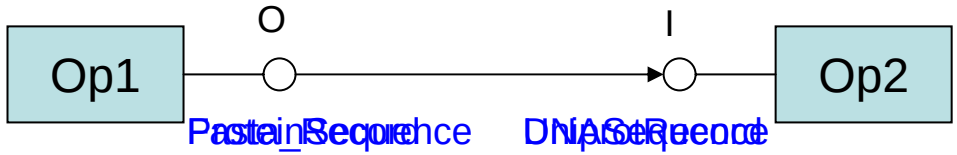
Type mismatch: In order to be compatible the data type of the output must be the same as or subtype of the data type required by the input parameter. The data link suffers from a type mismatch iff:

$$o.type \not\leq i.type$$

Cardinality mismatch: a particular kind of type mismatch. The data link suffers from a cardinality mismatch iff:

$$(o.type = ArrayOf\,(i.type))\ or$$
$$(i.type = ArrayOf\,(o.type))$$

Classes of mismatches



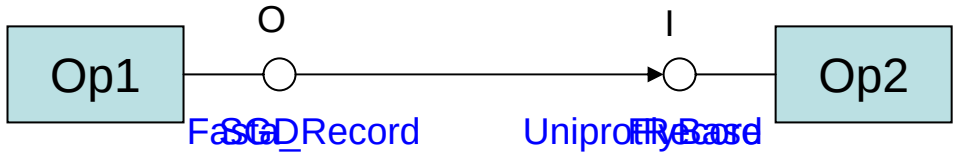
Domain mismatch: In order to be compatible the domain of the output must be the same as or subconcept of the domain of the subsequent input. The data link suffers from a domain mismatch iff:

$$domain(op1,o) \not\subseteq domain(op2,i)$$

Representation mismatch: refers to the difference in terms of format between the output and input. The data link suffers from a representation mismatch iff:

$$(domain(op1,o) \subseteq domain(op2,i)) \; and$$
$$(represent(op1,o) \neq represent(op2,i))$$

Classes of mismatches



Content mismatch: a particular kind of representation mismatch in which the formats conflict in terms of data scope. The data link suffers from a content mismatch iff:

$$contains(represent(op1,o),represent(op2,i)) = false$$

Extent mismatch: refers to the difference in terms of space of possible values between the output and input. The data link suffers from an extent mismatch iff:

$$coveredBy(extent(op1,o),extent(op2,i)) = false$$

Mapping

A mapping is used for transforming the data output by an operation onto the input of another operation

- Input/Output
- Domain preserving/ Non domain preserving
- Task

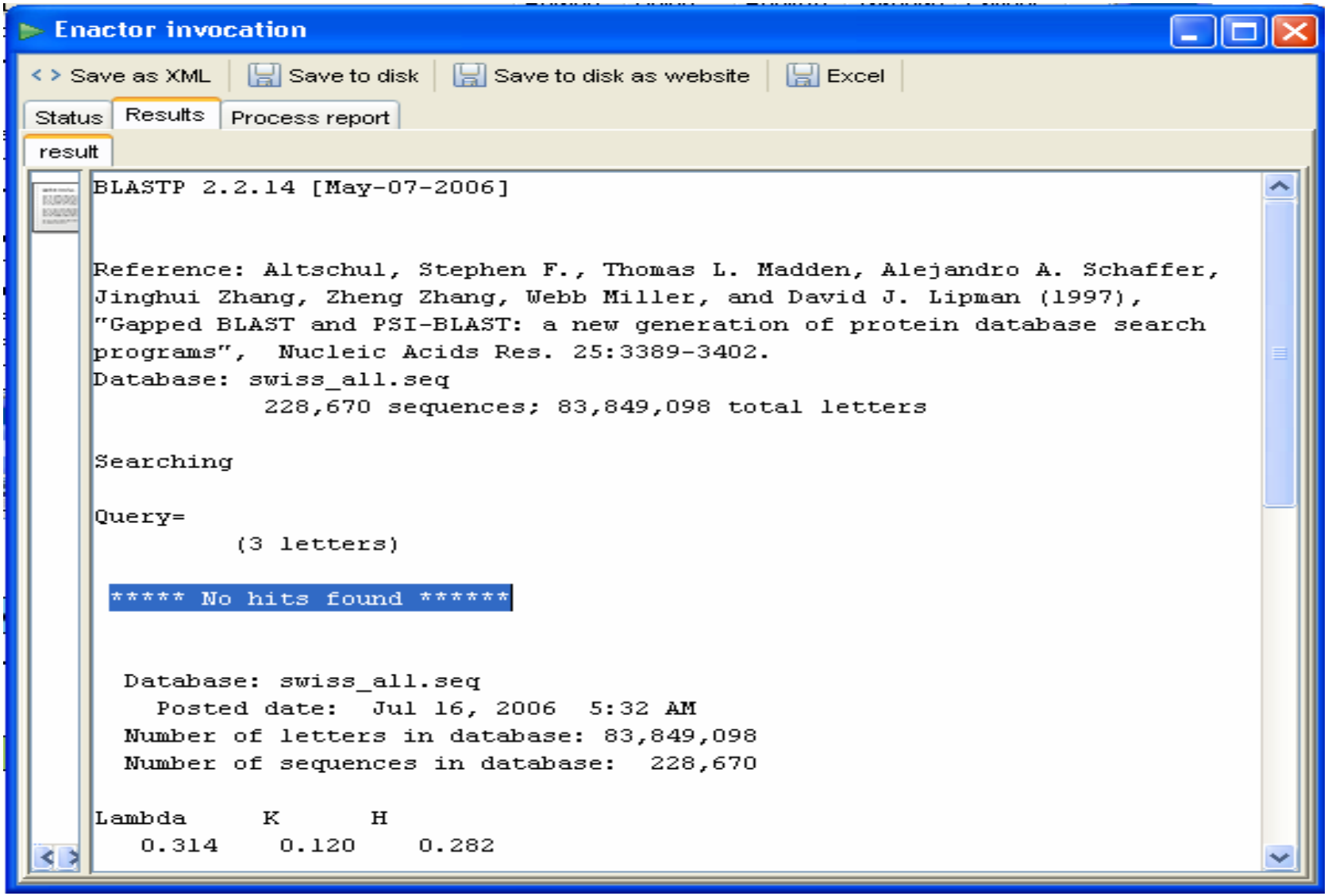
Evaluation

Workflow	Source	Mismatch
Value-Added Protein Identification	ISPIDER project	Domain and Content
Genome-focused identification	ISPIDER project	Type, Extent and Cardinality
Phylogenetic analysis	Hashmi *et al*	Domain and Representation
Arabidopsis genes prediction	myGrid project	Representation
Homology search	DDBJ	Representation
Gene Ontology Context	myGrid project	
Automatic refresh for Pride	ISPIDER project	Cardinality, Domain and Representation
Quality assessment workflow	Qurator project	
Genome annotation workflow	Pegasys project	Domain
Structure modeling workflow	myGrid project	Domain
Williams-Beuren Syndrome	myGrid project	Representation
Multiple alignment	EMBOSS	
Protein family analysis	REMORA	Domain and Representation

Conclusions

- A characterisation of mismatches
- A tool for automatically detecting mismatches and retrieving the mapping appropriate for their correction
- The developed tool has been used in practice
- Evaluation: the mismatches we characterised occur with different frequencies

Invalid results



Valid results