

# Mapping Polymorphism

**Ryan Wisnesky (Harvard)**

Mauricio Hernandez (IBM Almaden)

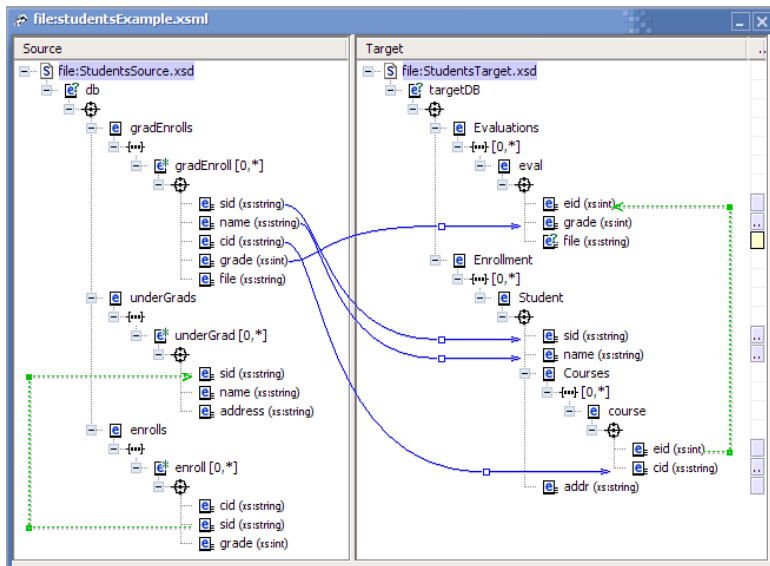
Lucian Popa (IBM Almaden)

ICDT 2010





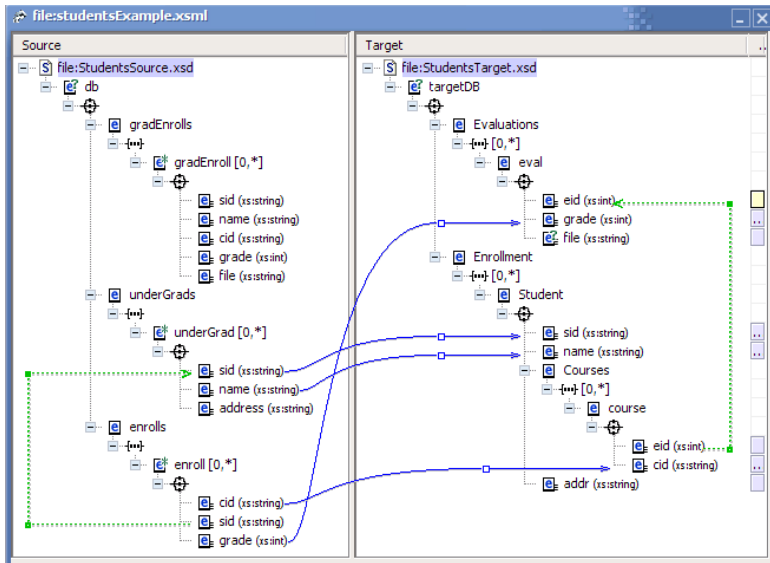
# A Schema Mapping in Clio



## A Schema Mapping in Clio

for  $g$  in  $db.gradEnrolls.gradEnroll$   
 $\Rightarrow$  exists  $s$  in  $targetDB.Enrollment.Student$ ,  
     $c$  in  $s.Courses.course$ ,  
     $e$  in  $targetDB.Evaluations.eval$   
where  $s.sid = g.sid \wedge s.name = g.name \wedge$   
     $c.cid = g.cid \wedge e.grade = g.grade \wedge$   
     $e.eid = c.eid$

# Nested, Alternating Quantifier Mappings



## Nested, Alternating Quantifier Mappings

for  $u$  in  $db.underGrads.underGrad$   
 $\Rightarrow$  exists  $s$  in  $targetDB.Enrollment.Student$   
where  $s.sid = u.sid \wedge s.name = u.name \wedge$   
    (for  $e$  in  $db.enrolls.enroll$   
    where  $e.sid = u.sid$   
     $\Rightarrow$  exists  $c$  in  $s.Courses.course,$   
           $e'$  in  $targetDB.Evaluations.eval$   
    where  $c.cid = e.cid \wedge e'.grade = e.grade \wedge$   
           $e'.eid = c.eid$ )

## Motivation: Mapping Re-use

- ▶ How can we adapt a mapping  $M : S \rightarrow T$  to new schema  $S'$  and  $T'$ ? Key questions:
  - ▶ When can we reuse  $M$  at  $S', T'$ , and
  - ▶ what does it mean when we do?
- ▶ Many approaches; for example: ask user for an  $M_1 : S' \rightarrow S$  and  $M_2 : T \rightarrow T'$ ; then compute  $M_2 \circ M \circ M_1$ .
- ▶ Our approach: reuse  $M$  directly when  $S', T' \leq S, T$ ; this is, when  $S', T'$  are subtypes (expansions) of  $S, T$ .

# Summary of Results

## Practice

---

1. Alice creates  $M : S \rightarrow T$ .
2. Clio computes principal typing  $\hat{S}, \hat{T}$  for  $M$ .
3. Bob wants to map  $S' \rightarrow T'$ , where  $S', T' \leq \hat{S}, \hat{T}$ .
4. Clio suggests using the mapping  $M : S' \rightarrow T'$ .



## Summary of Results

Practice	Theory
1. Alice creates $M : S \rightarrow T$ .	Type-checking ensures that $M$ is satisfiable.
2. Clio computes principal typing $\hat{S}, \hat{T}$ for $M$ .	Type-inference of $\hat{S}, \hat{T}$ is sound and complete.
3. Bob wants to map $S' \rightarrow T'$ , where $S', T' \leq \hat{S}, \hat{T}$ .	We coerce solutions to $M$ from $S', T'$ to $\hat{S}, \hat{T}$ .
4. Clio suggests using the mapping $M : S' \rightarrow T'$ .	Our semantics equates the meaning of $M : S \rightarrow T$ , $M : S' \rightarrow T'$ , and $M : \hat{S} \rightarrow \hat{T}$ .

# Outline

- ▶ **Formal Definitions of Schema, Mapping**
- ▶ Re-use walkthrough

# Our Nested Relational Model

$$\begin{aligned} \text{Row} & ::= - \mid \langle l : \text{Schema}, \text{Row} \rangle \\ \text{Schema} & ::= \text{ATOMIC } \mathcal{A} \mid \text{RCD } \text{Row} \mid \\ & \quad \text{SETRCD } \text{Row} \mid \text{SETCHC } \text{Row} \end{aligned}$$

- ▶ For this talk, we will assume rows are always well-formed.
- ▶ Order of labels in a row does not matter.
- ▶ We abbreviate  $\langle l_1 : t_1, \langle l_2 : t_2, - \rangle \rangle$  as  $\langle l_1 : t_1, l_2 : t_2 \rangle$ .
- ▶ Let  $\llbracket X \rrbracket$  denote the set of data instances corresponding to a schema  $X$ .

# Schema Roots

- ▶ A *context* provides types for the free variables of a mapping  $M$ , and an *environment* provides data instances.

## Definition (Context)

$$\Gamma ::= - \mid (v, \textit{Schema}); \Gamma$$

## Definition (Environment)

$$\Delta ::= - \mid (v, I); \Delta$$

- ▶ We write  $\Delta \in \llbracket \Gamma \rrbracket$  to indicate that for each  $(v, t) \in \Gamma$ , there is a corresponding  $(v, I) \in \Delta$  such that  $I \in \llbracket t \rrbracket$ .
- ▶ When  $M$  has two free variables  $src$  and  $dst$  corresponding to source data and data to be materialized, and  $(src, S); (dst, T)$  is a context for  $M$ , we write  $M : S \rightarrow T$ .

## NR Schema example

```
src, RCD (| students : SETRCD (|
    fullname : String,
    status : SETCHC (| teaching : String,
                    taking : String | | |)
|)
dst, RCD (| employees : SETRCD (| name : String,
    job : String,
    id : Int | |)
```

- ▶ DTD representation (ordered):

```
<!ELEMENT students (fullname , status)*>
<!ELEMENT status (teaching | taking)*> ...
<!ELEMENT employees (name, job, id)*> ...
```

## NR Data example

```
src, RCD (| students : SETRCD (|
      fullname : String,
      status : SETCHC (| teaching : String,
                       taking : String |) |)
dst, RCD (| employees : SETRCD (| name : String,
      job : String,
      id : Int |) |)
```

```
src, (students : {(fullname : John Doe,
                  status : {(teaching : CS100), (taking : CS200)}}
      ,
      (fullname : Mary Jane,
      status : {(taking : CS100), (taking : CS200)}}
      })
dst, (employees : {(name : John Doe, job : CS100, id : 1)})
```

# Compositional Mapping Expressions

$$\begin{aligned} Path & ::= v \mid Path.l \\ \diamond & ::= \underline{\text{for}} \mid \underline{\text{exists}} \\ \oplus & ::= \wedge \mid \Rightarrow \\ M & ::= \top \mid Path = Path \mid M \oplus M \mid \\ & \quad \diamond v \underline{\text{in}} Path . M \mid \\ & \quad \diamond v \underline{\text{of}} l \underline{\text{from}} Path . M \end{aligned}$$

- ▶ Our mappings are more expressive, and more compositional, than Clio mappings.

## Mapping Example

```
src, RCD ( students : SETRCD (
    fullname : String,
    status : SETCHC ( teaching : String,
                      taking : String ) ) )
dst, RCD ( employees : SETRCD ( name : String,
                                job : String,
                                id : Int ) )
```

```
for s in src.students .
  for t of teaching from s.status .
     $\top \Rightarrow$ 
    exists e in dst.employees .
       $e.name = s.fullname \wedge e.job = t$ 
```



# Standard Satisfaction Semantics

$$\frac{}{\Delta \models \top} \qquad \frac{\Delta \models m_1 \quad \Delta \models m_2}{\Delta \models m_1 \wedge m_2} \qquad \frac{\Delta \models m_1 \rightarrow \Delta \models m_2}{\Delta \models m_1 \Rightarrow m_2}$$

$$\frac{\Delta \models p_1 \rightsquigarrow I \quad \Delta \models p_2 \rightsquigarrow I}{\Delta \models p_1 = p_2} \qquad \frac{\Delta \models p \rightsquigarrow I \quad \forall i \in I, (v, i); \Delta \models m}{\Delta \models \text{for } v \text{ in } p. m}$$

$$\frac{\Delta \models p \rightsquigarrow I \quad \exists i \in I, (v, i); \Delta \models m}{\Delta \models \text{exists } v \text{ in } p. m}$$

$$\frac{\Delta \models p \rightsquigarrow I \quad \forall (l : i) \in I, (v, i); \Delta \models m}{\Delta \models \text{for } v \text{ of } l \text{ from } p. m}$$

$$\frac{\Delta \models p \rightsquigarrow I \quad \exists (l : i) \in I, (v, i); \Delta \models m}{\Delta \models \text{exists } v \text{ of } l \text{ from } p. m}$$

# Outline

- ▶ Formal Definitions of Schema, Mapping
- ▶ **Re-use walkthrough**

## Summary of Results

Practice	Theory
1. Alice creates $M : S \rightarrow T$ .	<b>Type-checking ensures that <math>M</math> is satisfiable.</b>
2. Clio computes principal typing $\hat{S}, \hat{T}$ for $M$ .	Type-inference of $\hat{S}, \hat{T}$ is sound and complete.
3. Bob wants to map $S' \rightarrow T'$ , where $S', T' \leq \hat{S}, \hat{T}$ .	We coerce solutions to $M$ from $S', T'$ to $\hat{S}, \hat{T}$ .
4. Clio suggests using the mapping $M : S' \rightarrow T'$ .	Our semantics equates the meaning of $M : S \rightarrow T$ , $M : S' \rightarrow T'$ , and $M : \hat{S} \rightarrow \hat{T}$ .

# Type-checking

$$\frac{(v, t) \in \Gamma}{\Gamma \vdash v :: t} \qquad \frac{\Gamma \vdash p :: \text{RCD } (\langle l : t, r \rangle)}{\Gamma \vdash p.l :: t}$$

$$\frac{\Gamma \vdash p_1 :: \text{ATOMIC } a \quad \Gamma \vdash p_2 :: \text{ATOMIC } a}{\Gamma \vdash p_1 = p_2} \qquad \frac{}{\Gamma \vdash \top}$$

$$\frac{\Gamma \vdash M_1 \quad \Gamma \vdash M_2}{\Gamma \vdash M_1 \oplus M_2} \qquad \frac{\Gamma \vdash p :: \text{SETRCD } r \quad (v, \text{RCD } r); \Gamma \vdash M}{\Gamma \vdash \diamond v \text{ in } p. M}$$

$$\frac{\Gamma \vdash p :: \text{SETCHC } (\langle l : t, r \rangle) \quad (v, t); \Gamma \vdash M}{\Gamma \vdash \diamond v \text{ of } l \text{ from } p. M}$$

# Type Safety

## Theorem

Suppose  $\Gamma \vdash M$ . Then  $M$  is satisfiable. That is, there exists a  $\Delta \in \llbracket \Gamma \rrbracket$  such that  $\Delta \models M$ .

- ▶ An example ill-typed, unsatisfiable mapping is:  
exists  $v$  in  $t$ .  $v = v.l$
- ▶ Given an  $M : S \rightarrow T$  and an instance  $I \in \llbracket S \rrbracket$ , typeability of  $M$  does not guarantee existence of a  $J \in \llbracket T \rrbracket$  such that  $I; J \models M$ .
  - ▶ Language of s-t tgds in [Fagin et al, TCS 2005] always admits solutions.
  - ▶ Language of nested mappings in [Fuxman et al, VLDB 2006] uses a complex syntactic check to guarantee solutions.

## Summary of Results

Practice	Theory
1. Alice creates $M : S \rightarrow T$ .	Type-checking ensures that $M$ is satisfiable.
<b>2. Clio computes principal typing <math>\hat{S}, \hat{T}</math> for <math>M</math>.</b>	<b>Type-inference of <math>\hat{S}, \hat{T}</math> is sound and complete.</b>
3. Bob wants to map $S' \rightarrow T'$ , where $S', T' \leq \hat{S}, \hat{T}$ .	We coerce solutions to $M$ from $S', T'$ to $\hat{S}, \hat{T}$ .
4. Clio suggests using the mapping $M : S' \rightarrow T'$ .	Our semantics equates the meaning of $M : S \rightarrow T$ , $M : S' \rightarrow T'$ , and $M : \hat{S} \rightarrow \hat{T}$ .

# Polymorphism

## Definition (Polymorphic NR Schema)

$$\begin{aligned} Row & ::= - \mid \langle Row, \mathcal{L} : Schema \rangle \mid \rho \\ Schema & ::= \text{ATOMIC } \alpha \mid \text{RCD } Row \mid \\ & \quad \text{SETRCD } Row \mid \text{SETCHC } Row \mid \sigma \end{aligned}$$

## Definition (Principal Typing)

$\hat{\Gamma}$  is a *principal typing* for  $M$  iff

1. for every substitution  $\phi$ , we have that  $\phi\hat{\Gamma} \vdash M$ .
2. for every  $\Gamma$  such that  $\Gamma \vdash M$ , there is some substitution  $\phi$  such that  $\phi\hat{\Gamma} = \Gamma$ .

## Principal Typing Example

$src$ , RCD (  $\rho_1$ , students : SETRCD (  $\rho_2$ ,  
                  fullname : ATOMIC  $\alpha_1$ , status : SETCHC (  $\rho_3$ ,  
                  teaching : ATOMIC  $\alpha_2$  ) ) ) )  
 $dst$ , RCD (  $\rho_4$ , employees : SETRCD (  $\rho_5$ , name : ATOMIC  $\alpha_1$ ,  
                  job : ATOMIC  $\alpha_2$  ) ) )

for  $s$  in  $src$ .students .  
  for  $t$  of teaching from  $s$ .status .  
     $\top \Rightarrow$   
    exists  $e$  in  $dst$ .employees .  
       $e$ .name =  $s$ .fullname  $\wedge$   $e$ .job =  $t$



# Type-inference

- ▶ The input to type inference is a mapping  $M$  and a context  $\Gamma$ , and the result is a substitution  $S$ . We write this as  $S\Gamma \Vdash M$ .
- ▶ Our particular algorithm is based on qualified types [Gaster and Jones, Hugs Haskell].
- ▶ Extend Hindley-Milner algorithm to account for permutation of rows:  $(l_1 : t_1, l_2 : t_2)$  must unify with  $(l_2 : t_2, l_1 : t_1)$ .

## Theorem (Soundness)

*For all  $\varphi\Gamma \Vdash M$ ,  $\varphi\Gamma \vdash M$ .*

## Theorem (Completeness)

*For all  $\varphi\Gamma \vdash M$ , there exists  $S$  and  $s$  such that  $S\Gamma \Vdash M$  and  $\varphi = s \circ S$ .*

## Summary of Results

Practice	Theory
1. Alice creates $M : S \rightarrow T$ .	Type-checking ensures that $M$ is satisfiable.
2. Clio computes principal typing $\hat{S}, \hat{T}$ for $M$ .	Type-inference of $\hat{S}, \hat{T}$ is sound and complete.
<b>3. Bob wants to map <math>S' \rightarrow T'</math>, where <math>S', T' \leq \hat{S}, \hat{T}</math>.</b>	<b>We coerce solutions to <math>M</math> from <math>S', T'</math> to <math>\hat{S}, \hat{T}</math>.</b>
4. Clio suggests using the mapping $M : S' \rightarrow T'$ .	Our semantics equates the meaning of $M : S \rightarrow T$ , $M : S' \rightarrow T'$ , and $M : \hat{S} \rightarrow \hat{T}$ .

# Structural Subtyping

$$\frac{}{\langle l : t, r \rangle \prec r} \quad \frac{t' < t \quad r' \preceq r}{\langle l : t', r' \rangle \prec \langle l : t, r \rangle} \quad \frac{r' \prec r}{\text{SETCHC } r' < \text{SETCHC } r}$$
$$\frac{r' \prec r}{\text{SETRCD } r' < \text{SETRCD } r} \quad \frac{r' \prec r}{\text{RCD } r' < \text{RCD } r}$$

*src*, RCD ( students : SETRCD ( fullname : String,  
status : SETCHC ( teaching : String, **taking : String** ) ) )

$\leq$

*src*, RCD ( students : SETRCD ( fullname : String,  
status : SETCHC ( teaching : String ) ) )

*dst*, RCD ( employees : SETRCD ( name : String, job : String, **id : Int** ) )

$\leq$

*dst*, RCD ( employees : SETRCD ( name : String, job : String ) )

# Subtyping and Semantics

- ▶ It is not the case that  $X \leq Y$  implies  $\llbracket X \rrbracket \subseteq \llbracket Y \rrbracket$ .
  - ▶ Example:  $\text{RCD } (\ell : \text{Int}) \not\leq \text{RCD } ()$
- ▶ But we can still relate subtyping to *erasure*.

## Definition (Erasure)

When  $\Gamma' \leq \Gamma$ , we can define an operation

$$\text{erase}(\Gamma' \leq \Gamma) : \llbracket \Gamma' \rrbracket \rightarrow \llbracket \Gamma \rrbracket$$

that removes data from instances in  $\llbracket \Gamma' \rrbracket$  so that they become instances in  $\llbracket \Gamma \rrbracket$ .

# Erasure

- ▶ The *erase* operation generalizes relational projection.
- ▶ For example,

$$\text{erase}(\text{SETRCD } \langle A : t_1, B : t_2 \rangle \leq \text{SETRCD } \langle A : t_1 \rangle)$$

is the same as projection from  $A, B$  to  $A$ .

## Theorem

*Suppose  $\Gamma \vdash M$  and  $\Gamma' \leq \Gamma$  and  $\Delta' \in \llbracket \Gamma' \rrbracket$ . Then  $\Delta' \models M$  if and only if  $\text{erase}(\Gamma' \leq \Gamma) (\Delta') \models M$ .*

## Summary of Results

Practice	Theory
1. Alice creates $M : S \rightarrow T$ .	Type-checking ensures that $M$ is satisfiable.
2. Clio computes principal typing $\hat{S}, \hat{T}$ for $M$ .	Type-inference of $\hat{S}, \hat{T}$ is sound and complete.
3. Bob wants to map $S' \rightarrow T'$ , where $S', T' \leq \hat{S}, \hat{T}$ .	We coerce solutions to $M$ from $S', T'$ to $\hat{S}, \hat{T}$ .
4. <b>Clio suggests using the mapping <math>M : S' \rightarrow T'</math>.</b>	<b>Our semantics equates the meaning of <math>M : S \rightarrow T</math>, <math>M : S' \rightarrow T'</math>, and <math>M : \hat{S} \rightarrow \hat{T}</math>.</b>

# Parametricity

## Definition

A mapping meaning function  $\llbracket \cdot \rrbracket$  is *parametric* if  $\Gamma \vdash M$  and  $\Gamma' \vdash M$  imply  $\llbracket (\Gamma', M) \rrbracket = \llbracket (\Gamma, M) \rrbracket$ .

- ▶ Parametric semantics are insensitive to irrelevant schema structure, and so are appropriate for re-use.
- ▶ A standard satisfaction-based semantics:

$$\llbracket (\Gamma, M) \rrbracket = \{ \Delta \mid \Delta \in \llbracket \Gamma \rrbracket \wedge \Delta \models M \}$$

is not parametric because as the schemas in  $\Gamma$  vary, so do the spaces of instances.

- ▶ Principal typings  $\hat{\Gamma}$  are unique, so a parametric semantics is:

$$\llbracket M \rrbracket = \{ \Delta \mid \Delta \in \llbracket \hat{\Gamma} \rrbracket \wedge \Delta \models M \}$$

## Conclusion

Practice	Theory
1. Alice creates $M : S \rightarrow T$ .	Type-checking ensures that $M$ is satisfiable.
2. Clio computes principal typing $\hat{S}, \hat{T}$ for $M$ .	Type-inference of $\hat{S}, \hat{T}$ is sound and complete.
3. Bob wants to map $S' \rightarrow T'$ , where $S', T' \leq \hat{S}, \hat{T}$ .	We coerce solutions to $M$ from $S', T'$ to $\hat{S}, \hat{T}$ .
4. Clio suggests using the mapping $M : S' \rightarrow T'$ .	Our semantics equates the meaning of $M : S \rightarrow T$ , $M : S' \rightarrow T'$ , and $M : \hat{S} \rightarrow \hat{T}$ .



## Related Work

- ▶ Nested Mappings (generation, compilation, etc)  
[Fuxman et al, VLDB 2006]
- ▶ Mapping semantics (cores, universal solutions, etc)  
[Fagin, TODS 2005]
- ▶ Operations over mappings (composition, inversion, etc)  
[Bernstein et al, VLDB 2006]  
[Fagin, TODS 2007]
- ▶ Mapping re-use (line matching, schema templates)  
[Madhavan et al, ICDE 2005]  
[Papotti et al, Data Knowl. Eng. 2009]

# Future Work

A types-based approach to mappings can also be used to study:

- ▶ Integration with programming languages
- ▶ Reuse under schema containment
- ▶ Recursion