

# View Generation and Optimisation in the AutoMed Data Integration Framework

AutoMed Technical Report 16, Version 3

Edgar Jasper<sup>1</sup>, Nerissa Tong<sup>2</sup>, Peter M<sup>c</sup>Brien<sup>2</sup>, and Alexandra Poulouvassilis<sup>1</sup>

<sup>1</sup> School of Computer Science and Information Systems, Birkbeck College,  
Univ. of London, {edgar,ap}@dcs.bbk.ac.uk

<sup>2</sup> Dept. of Computing, Imperial College, {nnyt98,pjm}@doc.ic.ac.uk

Monday 6<sup>th</sup> October 2003

## Abstract

This paper describes view generation and view optimisation in the AutoMed heterogeneous data integration framework. In AutoMed, schema integration is based on the use of reversible schema transformation sequences. We show how views can be generated from such sequences, for global-as-view (GAV), local-as-view (LAV) and GLAV query processing. We also present techniques for optimising these generated views, firstly by optimising the transformation sequences, and secondly by optimising the view definitions generated from them.

## 1 Introduction

Data integration is a process by which several databases, with associated **local schemas**, are integrated to form a single virtual database with an associated **global schema**. Up to now, most data integration approaches have been **global as view (GAV)** or **local as view (LAV)** [8]. In GAV, the constructs of a global schema are described as views over the local schemas. These view definitions are used to rewrite queries over a global schema into distributed queries over the local databases. Examples of the GAV approach are TSIMMIS [4], InterViso [22] and Garlic [21]. In LAV, the constructs of the local schemas are defined as views over the global schema, and processing queries over the global schema involves rewriting queries using views [9]. Examples of the LAV approach are IM [10] and Agora [12].

Both LAV and GAV lack a certain degree of expressiveness. GAV is unable to fully capture data integration semantics where a source schema construct can be defined by a non-reversible function over global schema constructs. For example, if source schema attribute **money** is the sum of global schema attributes **coins** and **notes**, neither **coins** nor **notes** in the global schema can be defined by views over the source schema. Thus a query on the global schema asking for the sum of **coins** and **notes** cannot be answered even though the answer (**money**) is present in the source schema). In LAV, the attribute **money** can be defined by a view as the sum of global schema attributes **coins** and **notes**. Conversely, reversing the presence of the attributes, so that **coins** and **notes** are in the local schema and **money** in the global schema, leads to a similar situation which GAV can express but LAV cannot.

GLAV [5] is a variation of LAV that allows the head of the view definition rules to contain conjunctions of relations from a source schema as a natural join, and is thus able to capture situations where a non-reversible function is a natural-join between attributes. In [11] GLAV was extended to allow any source schema query in the head of the rule, and hence is able to express the case where a single source schema is used to define the global schema constructs referenced in the body of the rule.

In [13] we presented a richer integration framework which is based on the use of reversible sequences of primitive schema transformations, called transformation **pathways**. In [17] we showed how these pathways incorporate the semantics of GAV rule definitions and LAV rule definitions, and hence we term our approach **both as view (BAV)**. We have implemented the BAV data integration approach within the AutoMed system (see <http://www.doc.ic.uk/automed>).

Since BAV integration is based on sequences of primitive schema transformations, it could be argued that the pathways resulting from BAV are likely to be more costly to reason with and process (*e.g.* for global query processing) than the corresponding LAV, GAV or GLAV view definitions would be. However, in Section 5 of this paper we show how BAV pathways are amenable to considerable simplification. Moreover, standard query optimisation techniques can be applied to the view definitions derived from BAV pathways, and we discuss these in Section 4 of this paper.

The outline of this paper is thus as follows. We begin with a brief review of the BAV integration approach in Section 2, give an example of its use, and compare it with the GAV, LAV and GLAV approaches. We then show how view definitions can be generated from BAV pathways for GAV, LAV or GLAV query processing, in Section 3. We then present techniques for optimising these generated views in Section 4. We finally present techniques for optimising the BAV pathways themselves in Section 5. Section 6 gives our concluding remarks and directions of further work.

## 2 The BAV Integration Approach

In previous work [20, 13] we have developed a general framework to support schema transformation and integration in heterogeneous database architectures. The framework consists of a low-level **hypergraph-based data model (HDM)** and a set of primitive schema transformations defined for this model. Higher-level data models and primitive schema transformations for them are defined in terms of this lower-level common data model.

In our framework, schemas are incrementally transformed by applying to them a sequence of primitive transformations  $t_1, \dots, t_r$ . Each primitive transformation  $t_i$  makes a ‘delta’ change to the schema by adding, deleting or renaming just one schema construct. Each add or delete transformation is accompanied by a query specifying the extent of the new or deleted construct in terms of the rest of the constructs in the schema. This query is expressed in a functional **intermediate query language, IQL**. All primitive transformations have an optional additional argument which specifies a constraint on the current schema extension that must hold if the transformation is to be applied. Constraints are also expressed as IQL queries.

A **composite transformation** is a sequence of primitive transformations. We term the composite transformation defined for transforming schema  $S_1$  to schema  $S_2$  a transformation **pathway**  $S_1 \rightarrow S_2$ . All source schemas, intermediate schemas and global schemas, and the

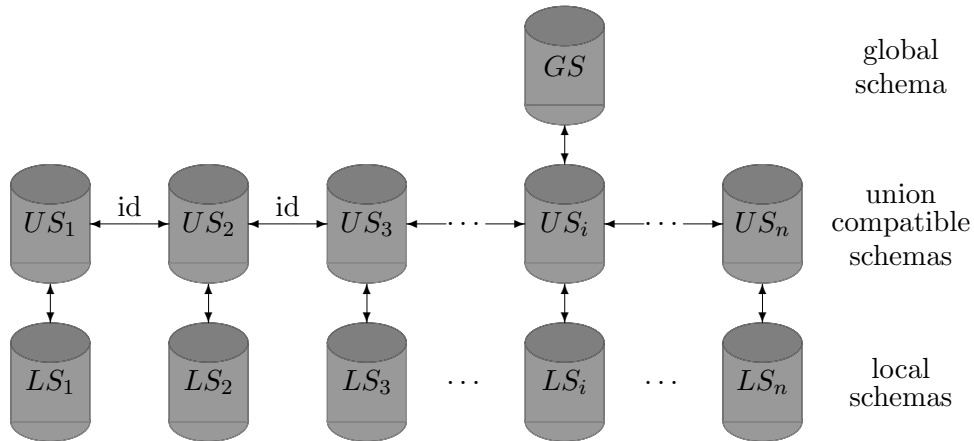


Figure 1: A general AutoMed Integration

pathways between them are stored in AutoMed’s metadata repository [1].

AutoMed supports a variety of methodologies for performing data integration and hence forming a **network** of pathways joining schemas together. For example, Figure 1 illustrates the integration of  $n$  local schemas,  $LS_1, \dots, LS_n$ , into a global schema  $GS$ . In order to integrate these  $n$  local schemas, each  $LS_i$  is first transformed into a “union” schema  $US_i$ . These  $n$  union schemas are syntactically identical, and this is asserted by creating a sequence of *id* transformation steps between each pair  $US_i$  and  $US_{i+1}$ , of the form *id* ( $US_i : c, US_{i+1} : c$ ) for each schema construct.

*id* is an additional type of primitive transformation, and the notation  $US_i : c$  is used to denote construct  $c$  appearing in schema  $US_i$ . These *id* transformations are generated automatically by the AutoMed software. An arbitrary one of the  $US_i$  can then be selected for further transformation into a global schema  $GS$ . This is where constructs sourced from different local schemas can be combined together by unions, joins, outer-joins etc.

There may be information within a  $US_i$  which is not semantically derivable from the corresponding  $LS_i$ . This is asserted by means of *extend* transformation steps within the pathway  $LS_i \rightarrow US_i$ . Conversely, not all of the information within a local schema  $LS_i$  need be transferred into  $US_i$  and this is asserted by means of *contract* transformation steps within  $LS_i \rightarrow US_i$ . These *extend* and *contract* transformations behave in the same way as *add* and *delete*, respectively, except that they indicate that only partial information can be derived about the new or deleted construct. Rather than a single query, they take a pair of queries which specify a lower and upper bound on the extent of the new or deleted construct. The lower bound query may be the constant *Void* if no lower bound can be specified, and the upper bound query may be the constant *Any* if no upper bound can be specified.

Each primitive transformation  $t$  has an **automatically derivable** reverse transformation  $\bar{t}$ . In particular, each *add* or *extend* transformation is reversed by a *delete* or *contract* transformation with the same arguments, and vice versa, while each *rename* or *id* transformation is reversed by another *rename* or *id* transformation with the two arguments swapped. This holds for the primitive transformations of **any** modelling language defined in AutoMed.

In [14] we described how our framework can be applied to different high-level modelling languages such as relational, ER and UML. The approach was extended to encompass XML

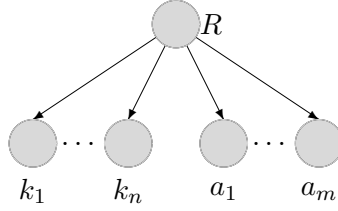


Figure 2: A simple relational data model

data sources in [15], formatted data files [6] and RDF [24]. For our purposes in the present paper, we assume that all schemas are specified in the very simple relational data model that we define below. However, we stress that the techniques that we describe here are equally applicable to any data modelling language supported by AutoMed.

Schemas in our simple relational model are constructed from primary key attributes, non-primary key attributes, and the relationships between them. Figure 2 illustrates the representation of a relation  $R$  with primary key attributes  $k_1, \dots, k_n$  and non-primary key attributes  $a_1, \dots, a_m$ . There is a one-one correspondence between this representation and the underlying HDM graph. In our simple relational model, there are two kinds of schema construct: **Rel** and **Att** (for simplicity, we ignore here the constraints present in a relational schema but see [13] for an encoding of a richer relational data model).

The extent of a **Rel** construct  $\langle\langle R \rangle\rangle$  is the projection of the relation  $R$  onto its primary key attributes  $k_1, \dots, k_n$ . The extent of each **Att** construct  $\langle\langle R, a \rangle\rangle$  where  $a$  is an attribute (key or non-key) is the projection of relation  $R$  onto  $k_1, \dots, k_n, a$ . For example, a relation `student(id,sex,dname)` would be modelled by a **Rel** construct  $\langle\langle \text{student} \rangle\rangle$ , and three **Att** constructs  $\langle\langle \text{student}, \text{id} \rangle\rangle$ ,  $\langle\langle \text{student}, \text{sex} \rangle\rangle$  and  $\langle\langle \text{student}, \text{dname} \rangle\rangle$ <sup>1</sup>.

Once the constructs of modelling language  $\mathcal{M}$  have been defined in terms of the HDM (via the API of AutoMed’s metadata repository [1]), a set of primitive schema transformations for  $\mathcal{M}$  are automatically available. For the simple relational model above, these would be as follows:

- **addRel**( $\langle\langle R \rangle\rangle, q$ ) adds to the schema a new relation  $R$ . The query  $q$  specifies the set of primary key values in the extent of  $R$  in terms of the already existing schema constructs.
- **addAtt**( $\langle\langle R, a \rangle\rangle, q$ ) adds to the schema an attribute  $a$  (key or non-key) for relation  $R$ . The query  $q$  specifies the extent of the binary relationship between the primary key attribute(s) of  $R$  and this new attribute  $a$  in terms of the already existing schema constructs.
- **deleteRel**( $\langle\langle R \rangle\rangle, q$ ) deletes from the schema the relation  $R$  (provided all its attributes have first been deleted). The query  $q$  specifies how the extent of  $R$  can be restored from the remaining schema constructs.

<sup>1</sup>We note that there is some redundancy of representation here, since the extent of each primary key attribute  $k_i$  appears both within the extent of  $\langle\langle R \rangle\rangle$  and within the extent of  $\langle\langle R, k_i \rangle\rangle$ . We have adopted this representation for two main reasons. Firstly, it allows for easier integration of relational schemas with object-oriented schemas, in that  $\langle\langle R \rangle\rangle$  corresponds to an OO class and each  $\langle\langle R, a \rangle\rangle$  to an OO attribute. Secondly, this uniformity of treatment for all attributes allows for easier translation of IQL queries into SQL queries on local relational data sources.

<p>LS<sub>1</sub> staff(<u>id</u>,name,dname) male(<i>id</i>) female(<i>id</i>)</p>	<p>US<sub>i</sub> university(<u>uname</u>) campus(<u>cmname</u>,<i>uname</i>) dept(<u>dname</u>,street,<i>cmname</i>) degree(<u>dcode</u>,title,dtype,<i>dname</i>) staff(<u>id</u>,name,sex,<i>dname</i>) student(<u>id</u>,name,sex) enrolled(<i>id</i>,<u>from</u>,to,<i>dcode</i>)</p>
<p>LS<sub>2</sub> university(uname) campus(<u>cmname</u>,<i>uname</i>) dept(<u>deptname</u>,<i>cmname</i>) degree(<u>dcode</u>,title,dtype,<i>deptname</i>)</p>	<p>GS university(<u>uname</u>) campus(<u>cmname</u>,<i>uname</i>) dept(<u>dname</u>,street,<i>cmname</i>) degree(<u>dcode</u>,title,dtype,<i>dname</i>) person(<u>id</u>,name,sex,<i>dname</i>#) enrolled(<i>id</i>,<u>from</u>,to,<i>dcode</i>)</p>
<p>LS<sub>3</sub> student(<u>id</u>,name,sex) enrolled(<i>id</i>,<u>from</u>,to,<i>dcode</i>) degree(<u>dcode</u>)</p>	
<p>LS<sub>4</sub> university(uname) college(<u>cname</u>,<i>uname</i>) dept(<u>dname</u>,street,<i>cname</i>)</p>	

Figure 3: Example schemas

- `deleteAtt( $\langle\langle R, a \rangle\rangle, q$ )` deletes from the schema attribute  $a$  of relation  $R$ . The query  $q$  specifies how the extent of the binary relationship between the primary key attribute(s) of  $R$  and  $a$  can be restored from the remaining schema constructs.
- `renameRel( $\langle\langle R \rangle\rangle, \langle\langle R' \rangle\rangle$ )` renames the relation  $R$  to  $R'$  in the schema.
- `renameAtt( $\langle\langle R, a \rangle\rangle, \langle\langle R, a' \rangle\rangle$ )` renames the attribute  $a$  of  $R$  to  $a'$ .

There is also a set of `extendRel`, `extendAtt`, `contractRel` and `contractAtt` primitive transformations.

## 2.1 An Example Integration

Figure 3 gives some specific schemas to illustrate the integration approach of Figure 1. Primary key attributes are underlined, foreign key attributes are in italics and nullable attributes are suffixed by #.

In Example 1, transformations  $t_1$ – $t_5$  use a composite transformation `extendTable` to state that the tables `student`, `enrolled`, `university`, `campus` and `degree` in  $US_1$  cannot be derived from  $LS_1$ . The definition of `extendTable` is:

$$\begin{aligned} \text{extendTable}(\langle\langle R, a_1, \dots, a_n \rangle\rangle) &= \text{extendRel}(\langle\langle R \rangle\rangle, \text{Void}, \text{Any}) \\ &\quad \text{extendAtt}(\langle\langle R, a_1 \rangle\rangle, \text{Void}, \text{Any}) \\ &\quad \vdots \\ &\quad \text{extendAtt}(\langle\langle R, a_n \rangle\rangle, \text{Void}, \text{Any}) \end{aligned}$$

Then transformations  $t_6$ – $t_9$  use the `dname` attribute of `staff` to derive the `dept` table in  $US_1$ , and use `extend` transformations for the two attributes `street` and `uname` that cannot be derived from  $LS_1$ . Finally, in  $t_{10}$ – $t_{14}$  the `male` and `female` relations of  $LS_1$  are restructured into the single `sex` attribute of `staff`.

The queries accompanying the **add** and **delete** transformations are expressed in our IQL intermediate query language. In IQL,  $++$  is the bag union operator and the construct  $[e \mid Q_1; \dots Q_n]$  is a **comprehension** [2]. The expressions  $Q_1$  to  $Q_n$  are termed **qualifiers**, each qualifier being either a **filter** or a **generator**. A filter is a boolean-valued expression. A generator has syntax  $p \leftarrow c$  where  $p$  is a **pattern** and  $c$  is a bag-valued expression. In IQL, the patterns  $p$  are restricted to be single variables or tuples of variables.

**Example 1 Pathway  $LS_1 \rightarrow US_1$**

```

t1  extendTable(⟨⟨student, id, name, sex⟩⟩)
t2  extendTable(⟨⟨university, unname⟩⟩)
t3  extendTable(⟨⟨campus, cmname, unname⟩⟩)
t4  extendTable(⟨⟨degree, dcode, title, dtype, dname⟩⟩)
t5  extendTable(⟨⟨enrolled, id, from, to, dcode⟩⟩)
t6  addRel(⟨⟨dept⟩⟩, [x | (y, x) ← ⟨⟨staff, dname⟩⟩])
t7  addAtt(⟨⟨dept, dname⟩⟩, [(x, x) | x ← ⟨⟨dept⟩⟩])
t8  extendAtt(⟨⟨dept, street⟩⟩, Void, Any)
t9  extendAtt(⟨⟨dept, unname⟩⟩, Void, Any)
t10 addAtt(⟨⟨staff, sex⟩⟩, [(x, 'M') | x ← ⟨⟨male⟩⟩] ++ [(x, 'F') | x ← ⟨⟨female⟩⟩])
t11 deleteAtt(⟨⟨male, id⟩⟩, [(x, x) | x ← ⟨⟨male⟩⟩])
t12 deleteRel(⟨⟨male⟩⟩, [x | (x, 'M') ← ⟨⟨staff, sex⟩⟩])
t13 deleteAtt(⟨⟨female, id⟩⟩, [(x, x) | x ← ⟨⟨female⟩⟩])
t14 deleteRel(⟨⟨female⟩⟩, [x | (x, 'F') ← ⟨⟨staff, sex⟩⟩])

```

The pathway  $LS_2 \rightarrow US_2$  contains extend steps to add the missing **staff**, **student**, and **enrolled** tables. It then renames **deptname**, and adds the missing attributes of **dept**:

**Example 2 Pathway  $LS_2 \rightarrow US_2$**

```

t15  extendTable(⟨⟨student, id, name, sex⟩⟩)
t16  extendTable(⟨⟨staff, id, name, sex, dname⟩⟩)
t17  extendTable(⟨⟨enrolled, id, from, to, dcode⟩⟩)
t18  renameAtt(⟨⟨dept, deptname⟩⟩, ⟨⟨dept, dname⟩⟩)
t19  renameAtt(⟨⟨degree, deptname⟩⟩, ⟨⟨degree, dname⟩⟩)
t20  extendAtt(⟨⟨dept, street⟩⟩, Void, Any)
t21  extendAtt(⟨⟨dept, unname⟩⟩, Void, Any)

```

The pathway  $LS_3 \rightarrow US_3$  contains a sequence of **extend** steps for its missing information. The pathway  $LS_4 \rightarrow US_4$  creates in  $t_{22}$  a new attribute  $\langle\langle\text{dept, unname}\rangle\rangle$  by joining the **dept** and **college** relations, and then deletes in  $t_{23}$ – $t_{25}$  the **college** table that can be recovered from the remaining  $\langle\langle\text{dept, cname}\rangle\rangle$  attribute. Transformation  $t_{26}$  is unable to put any restriction on the values of  $\langle\langle\text{dept, cname}\rangle\rangle$ , since that association cannot be recovered from the global schema. Transformations  $t_{27}$ – $t_{31}$  then perform the logical inverse of  $t_{22}$ – $t_{26}$  to partially extract the **campus** table from the direct association between departments and universities represented by  $\langle\langle\text{dept, unname}\rangle\rangle$ .

**Example 3 Pathway  $LS_4 \rightarrow US_4$** 

```

t22  addAtt(⟨⟨dept, uname⟩⟩, [(x, y) | (x, z) ← ⟨⟨dept, cname⟩⟩; (z, y) ← ⟨⟨college, uname⟩⟩])
t23  deleteAtt(⟨⟨college, uname⟩⟩, [(x, y) | (z, x) ← ⟨⟨dept, cname⟩⟩; (z, y) ← ⟨⟨dept, uname⟩⟩])
t24  deleteAtt(⟨⟨college, cname⟩⟩, [(x, x) | x ← ⟨⟨college⟩⟩])
t25  deleteRel(⟨⟨college⟩⟩, [y | (x, y) ← ⟨⟨dept, cname⟩⟩])
t26  contractAtt(⟨⟨dept, cname⟩⟩, Void, Any)
t27  extendAtt(⟨⟨dept, cmname⟩⟩, Void, Any)
t28  addRel(⟨⟨campus⟩⟩, [y | (x, y) ← ⟨⟨dept, cmname⟩⟩])
t29  addAtt(⟨⟨campus, cmname⟩⟩, [(x, x) | x ← ⟨⟨campus⟩⟩])
t30  addAtt(⟨⟨campus, uname⟩⟩, [(x, y) | (z, x) ← ⟨⟨dept, cmname⟩⟩; (z, y) ← ⟨⟨dept, uname⟩⟩])
t31  delAtt(⟨⟨dept, uname⟩⟩, [(x, y) | (x, z) ← ⟨⟨dept, cmname⟩⟩; (z, y) ← ⟨⟨campus, uname⟩⟩])
t32  extendTable(⟨⟨student, id, name, sex⟩⟩)
t33  extendTable(⟨⟨staff, id, name, sex, dname⟩⟩)
t34  extendTable(⟨⟨enrolled, id, from, to, dcode⟩⟩)

```

Finally, we list in Example 4 the pathway from the union schema  $US_1$  to the global schema  $GS$ . The pathway from  $US_2$ ,  $US_3$  or  $US_4$  would be identical.

**Example 4 Pathway  $US_1 \rightarrow GS$** 

```

t35  addRel(⟨⟨person⟩⟩, ⟨⟨staff⟩⟩ ++ [x | x ← ⟨⟨student⟩⟩; not (member x ⟨⟨staff⟩⟩)])
t36  addAtt(⟨⟨person, id⟩⟩, ⟨⟨staff, id⟩⟩ ++ [(x, y) | (x, y) ← ⟨⟨student, id⟩⟩; not (member x ⟨⟨staff⟩⟩)])
t37  addAtt(⟨⟨person, name⟩⟩, ⟨⟨staff, name⟩⟩ ++
        [(x, y) | (x, y) ← ⟨⟨student, name⟩⟩; not (member x ⟨⟨staff⟩⟩)])
t38  addAtt(⟨⟨person, sex⟩⟩, ⟨⟨staff, sex⟩⟩ ++ [(x, y) | (x, y) ← ⟨⟨student, sex⟩⟩; not (member x ⟨⟨staff⟩⟩)])
t39  addAtt(⟨⟨person, dname⟩⟩, ⟨⟨staff, dname⟩⟩)
t40  contractAtt(⟨⟨student, id⟩⟩, [(x, y) | (x, y) ← ⟨⟨person, id⟩⟩; not (member x ⟨⟨staff⟩⟩)],
               [(x, y) | (x, y) ← ⟨⟨person, id⟩⟩])
t41  contractAtt(⟨⟨student, name⟩⟩, [(x, y) | (x, y) ← ⟨⟨person, name⟩⟩; not (member x ⟨⟨staff⟩⟩)],
               [(x, y) | (x, y) ← ⟨⟨person, name⟩⟩])
t42  contractAtt(⟨⟨student, sex⟩⟩, [(x, y) | (x, y) ← ⟨⟨person, sex⟩⟩; not (member x ⟨⟨staff⟩⟩)],
               [(x, y) | (x, y) ← ⟨⟨person, sex⟩⟩])
t43  contractRel(⟨⟨student⟩⟩, [x | x ← ⟨⟨person⟩⟩; not (member x ⟨⟨staff⟩⟩)],
               [x | x ← ⟨⟨person⟩⟩])
t44  deleteAtt(⟨⟨staff, id⟩⟩, [(x, y) | (x, y) ← ⟨⟨person, id⟩⟩; member x ⟨⟨staff⟩⟩])
t45  deleteAtt(⟨⟨staff, name⟩⟩, [(x, y) | (x, y) ← ⟨⟨person, name⟩⟩; member x ⟨⟨staff⟩⟩])
t46  deleteAtt(⟨⟨staff, sex⟩⟩, [(x, y) | (x, y) ← ⟨⟨person, sex⟩⟩; member x ⟨⟨staff⟩⟩])
t47  deleteAtt(⟨⟨staff, dname⟩⟩, ⟨⟨person, dname⟩⟩)
t48  deleteRel(⟨⟨staff⟩⟩, [x | (x, y) ← ⟨⟨person, dname⟩⟩])

```

We assume in this example integration that a person may be both a member of staff and a student. For such people, their information in the `staff` table is preferred for propagation to the global `person` table in steps  $t_{35}$ – $t_{38}$  above. Thus, there is not sufficient information in the global schema to totally derive the `student` table, and only `contract` statements can be given in steps  $t_{40}$ – $t_{43}$ , where as a lower bound we know all persons not in the `staff` table are students, but as an upper bound know that all persons could be in `student` (if it were the case that all staff members were former students). Conversely, there is sufficient information to totally derive the `staff` table.

## 2.2 Comparison of BAV with GAV, LAV and GLAV

We see from the above example that the **add** and **extend** steps in the transformation pathways from the local schemas to the global schema correspond to GAV, since it is these steps that are incrementally defining global constructs in terms of local ones. Similarly, it is the **delete** and **contract** steps in the transformation pathways from the local schemas to the global schema that correspond to LAV, since it is these steps that are incrementally defining local constructs in terms of global ones. We will see in Section 3 how these pathways can be traversed to derive GAV and LAV views.

If a GAV view is derived from solely **add** steps it will be *exact* in the terminology of [7]. If, in addition, it is derived from one or more **extend** steps using their lower-bound (upper-bound) queries, then the GAV view will be *sound* (*complete*) in the terminology of [7]. Similarly, if a LAV view is derived solely from **delete** steps it will be exact. If, in addition, it is derived from one or more **contract** steps using their lower-bound (upper-bound) queries, then the LAV view will be *complete* (*sound*) in the terminology of [7]. For example, in pathway  $US_1 \rightarrow GS$  above, we could enhance  $t_{43}$  above to:

`contractRel(⟨⟨student⟩⟩, [x | x ← ⟨⟨person⟩⟩; not (member x ⟨⟨staff⟩⟩)], ⟨⟨person⟩⟩)`

asserting that  $\langle\langle\text{student}\rangle\rangle$  contains the set of people who are not staff (completeness) and is contained by the whole set of people (soundness).

As we discussed in the introduction, BAV is a more expressive data integration language than LAV, GAV or GLAV, since it allows for the expression of mappings in both directions, and since it is not limited on how many source schemas are associated by a mapping. Indeed, in the context of peer-to-peer integration, [3] has suggested using GLAV rules in both directions in a similar manner to BAV, in order to overcome weaknesses of using GLAV alone.

As discussed in [16, 17], a further advantage of BAV over GAV and LAV is that it readily supports the evolution of both global and local schemas, by allowing pathways and schemas to be incrementally modified as opposed to having to be regenerated.

A further difference between BAV and GAV, LAV or GLAV (including the approach of using GLAV in each direction of [3]) is that statements about the relationships between global and local schemas are made at a finer level of detail, *i.e.* at the level of individual attributes as opposed to entire tables. So we can assert exact knowledge about some attributes of a table, and sound or complete knowledge about other attributes. We are also able to introduce intermediate constructs in the mapping, such as in  $LS_4 \rightarrow US_4$ .

## 3 Generating Views

We now present a general technique for generating GAV, LAV and GLAV view definitions from a BAV pathway. This ability to generate any of these kinds of view definitions from a single BAV pathway means that we can select a query processing technique that can vary between queries as appropriate.

To define a construct  $c$  in  $S_x$  in terms of the constructs in schema  $S_y$ , we consider in turn the transformations of  $S_x \rightarrow S_y$ . The only transformations that are significant are those that delete, contract or rename a construct<sup>2</sup>. These transformations are significant because the current view definitions may query constructs that no longer exist after such a transformation. Each of these types of transformation is handled as follows if it is encountered during the

---

<sup>2</sup>Note that this is equivalent to considering the **add**, **extend** and **rename** steps in the reverse  $S_y \rightarrow S_x$



traversal of  $S_x \rightarrow S_y$ :

- **delete**: This has an associated query which shows how to reconstruct the extent of the construct being deleted. Any occurrence of the deleted construct within the current view definitions is replaced by this query.
- **contract**: Any occurrence of the contracted construct within the current view definitions is replaced by either the lower-bound or the upper-bound query accompanying this transformation step, depending on whether sound or complete views are required.
- **rename**: All references to the old construct in the current view definitions are replaced by references to the new construct.

### 3.1 Generating GAV Views

To generate the set of GAV views for a global schema, the pathways from it to each local schema are retrieved from AutoMed’s metadata repository. For some part of the start of their length these pathways may be the same, as may be seen from the tree structure of Figure 1. Each node of this transformations tree is a schema (global, intermediate or local) linked to its neighbours by a single transformation step. View definitions for each global schema construct are derived by traversing the tree from top to bottom. Initially, each construct’s view definition is just the construct itself. Each node in the tree is then visited in a downwards direction, and **delete**, **contract** and **rename** transformations are handled as described above. In particular, if a **contract** transformation step is encountered, any occurrence of the contracted construct within the current GAV view definitions is replaced by the lower-bound query accompanying this transformation step (so that sound GAV views will be generated).

At some points the tree may branch. When this happens, constructs of the parent schema are semantically identical to constructs that have the same scheme within the child schemas. The possibility of using all paths is retained within the view definitions by replacing each construct of the parent schema by a disjunction (OR) of the corresponding constructs in the child schemas.

The tree is traversed in this fashion from the root to the leaves until all the nodes are visited. The resulting view definitions are the GAV definitions for the global schema constructs over the local schemas. Referring again to the example of Section 2.1, consider the construct  $GS: \langle\langle \text{person}, \text{sex} \rangle\rangle$  in the global schema. The pathway  $GS \rightarrow US_1$  would be processed first (*i.e.* the reverse of the pathway  $US_1 \rightarrow GS$  listed in Section 2.1). The only significant transformation is  $\bar{t}_{38}$  that deletes  $\langle\langle \text{person}, \text{sex} \rangle\rangle$ , resulting in an intermediate view definition:  $GS: \langle\langle \text{person}, \text{sex} \rangle\rangle :-$

$US_1: \langle\langle \text{staff}, \text{sex} \rangle\rangle ++ [(x, y) \mid (x, y) \leftarrow US_1: \langle\langle \text{student}, \text{sex} \rangle\rangle; \text{not} (\text{member } x \text{ } US_1: \langle\langle \text{staff} \rangle\rangle)]$

at one copy,  $US_1$ , of the four union schemas. Traversing the pathways  $US_1 \rightarrow LS_1$  and  $US_1 \rightarrow US_2$ , we replace the body of the view definition with:

$( [(x, 'M') \mid x \leftarrow LS_1: \langle\langle \text{male} \rangle\rangle] ++ [(x, 'F') \mid x \leftarrow LS_1: \langle\langle \text{female} \rangle\rangle] \text{ OR } US_2: \langle\langle \text{staff}, \text{sex} \rangle\rangle)$   
 $++ ( [(x, y) \mid (x, y) \leftarrow \text{Void OR } US_2: \langle\langle \text{student}, \text{sex} \rangle\rangle;$   
 $\text{not} (\text{member } x \text{ } (LS_1: \langle\langle \text{staff} \rangle\rangle \text{ OR } US_2: \langle\langle \text{staff} \rangle\rangle))])$

Traversing next  $US_2 \rightarrow LS_2$  and  $US_2 \rightarrow US_3$ , we get:

$( [(x, 'M') \mid x \leftarrow LS_1: \langle\langle \text{male} \rangle\rangle] ++ [(x, 'F') \mid x \leftarrow LS_1: \langle\langle \text{female} \rangle\rangle]) \text{ OR}$   
 $\text{Void OR } US_3: \langle\langle \text{staff}, \text{sex} \rangle\rangle)$   
 $++ ( [(x, y) \mid (x, y) \leftarrow \text{Void OR Void OR } US_3: \langle\langle \text{student}, \text{sex} \rangle\rangle;$   
 $\text{not} (\text{member } x \text{ } (LS_1: \langle\langle \text{staff} \rangle\rangle \text{ OR Void OR } US_3: \langle\langle \text{staff} \rangle\rangle))])$

Continuing with  $US_3 \rightarrow LS_3$  and  $US_3 \rightarrow US_4$ , we get:

```
( [(x, 'M') | x ← LS1:⟨⟨male⟩⟩] ++ [(x, 'F') | x ← LS1:⟨⟨female⟩⟩]) OR
Void OR Void OR US4:⟨⟨staff, sex⟩⟩
++ ( [(x, y) | (x, y) ← Void OR Void OR LS3:⟨⟨student, sex⟩⟩ OR US4:⟨⟨student, sex⟩⟩;
not (member x (LS1:⟨⟨staff⟩⟩ OR Void OR Void OR US4:⟨⟨staff⟩⟩))])
```

Traversing finally  $US_4 \rightarrow LS_4$ , we obtain the view definition:

```
GS:⟨⟨person, sex⟩⟩ :-
( [(x, 'M') | x ← LS1:⟨⟨male⟩⟩] ++ [(x, 'F') | x ← LS1:⟨⟨female⟩⟩]) OR
Void OR Void OR Void
++ ( [(x, y) | (x, y) ← Void OR Void OR LS3:⟨⟨student, sex⟩⟩ OR Void;
not (member x (LS1:⟨⟨staff⟩⟩ OR Void OR Void OR Void))])
```

Section 4 will justify how this view definition can be simplified to:

```
GS:⟨⟨person, sex⟩⟩ :-
( [(x, 'M') | x ← LS1:⟨⟨male⟩⟩] ++ [(x, 'F') | x ← LS1:⟨⟨female⟩⟩])
++ ( [(x, y) | (x, y) ← LS3:⟨⟨student, sex⟩⟩; not (member x (LS1:⟨⟨staff⟩⟩))])
```

Such view derivations can be substituted into any query posed on a global schema in order to obtain an equivalent query distributed over the local schemas — this is the GAV approach to global query processing, which is what the AutoMed implementation currently supports.

### 3.2 Generating LAV Views

LAV views are derived similarly: the pathway from a local schema to the global schema is again retrieved from the metadata repository and is processed as above to derive the view definitions, except that it is the local schema end of the pathway that is now taken as the root of the tree. The derivation of LAV views is simpler because there is now only a single pathway being processed, with no branching. Also, if a contract transformation step is encountered, any occurrence of the contracted construct within the current LAV view definitions is replaced by the upper-bound query accompanying this transformation step (so that sound LAV views will be generated).

For example, to generate a LAV definition of  $LS_1 : \langle \langle \text{male} \rangle \rangle$ , we inspect the pathway  $t_1, \dots, t_{14}, t_{35}, \dots, t_{48}$ . The transformation  $t_{12}$  deletes  $\langle \langle \text{male} \rangle \rangle$ , and therefore we have an intermediate view definition on  $US_1$ :

```
LS1:⟨⟨male⟩⟩ :- [x | (x, 'M') ← US1:⟨⟨staff, sex⟩⟩]
```

Then  $\langle \langle \text{staff, sex} \rangle \rangle$  construct is deleted by  $t_{46}$ , which substitutes  $(x, y) \leftarrow \langle \langle \text{staff, sex} \rangle \rangle$  with  $(x, y) \leftarrow \langle \langle \text{person, sex} \rangle \rangle$ ;  $\text{member } x \langle \langle \text{staff} \rangle \rangle$ , and the  $\langle \langle \text{staff} \rangle \rangle$  construct in this query is deleted by  $t_{40}$  giving a final LAV rule:

```
LS1:⟨⟨male⟩⟩ :- [x | (x, 'M') ← GS:⟨⟨person, sex⟩⟩; member x [x | (x, y) ← GS:⟨⟨person, dname⟩⟩]]
```

### 3.3 Generating GLAV Views

First, it should be noted that GLAV view definitions will include all the LAV view definitions, and all the GAV view definitions where the body of the rule is a query that matches the conditions required for the GLAV query processing system in use (which in [11] would be queries over a single source). In addition, we inspect now all the add and extend transformations of the pathway that would be ignored by LAV view generation, and for each one use the query to form the head of a new GLAV rule.

For example, in  $LS_4 \rightarrow US_4$ , the query in  $t_{22}$  gives a new view rule head:

$[(x, y) \mid (x, z) \leftarrow \text{LS}_4 : \langle\langle \text{dept}, \text{cname} \rangle\rangle; (z, y) \leftarrow \text{LS}_4 : \langle\langle \text{college}, \text{uname} \rangle\rangle]$  which is defined by  $\langle\langle \text{dept}, \text{uname} \rangle\rangle$  at this stage. We then use our standard algorithm on construct  $\langle\langle \text{dept}, \text{uname} \rangle\rangle$ , detect that it is deleted in  $t_{31}$ , and hence replace it with the query from  $t_{31}$  to result in the GLAV rule:

$[(x, y) \mid (x, z) \leftarrow \text{LS}_4 : \langle\langle \text{dept}, \text{cname} \rangle\rangle; (z, y) \leftarrow \text{LS}_4 : \langle\langle \text{college}, \text{uname} \rangle\rangle] :-$   
 $[(x, z) \mid \text{GS} : \langle\langle \text{dept}, \text{cmname} \rangle\rangle; (z, y) \leftarrow \text{GS} : \langle\langle \text{campus}, \text{uname} \rangle\rangle]$

Note however that the BAV integration would still hold if  $\text{LS}_4$  were fragmented, with campus and departments held on separate sources, whereas GLAV would cease to be valid in this situation.

## 4 Optimising the Generated Views

The view definitions generated by the process described above can be simplified after they have been derived. This saves later work for the query optimiser when these definitions are substituted into specific global queries for query processing. It also means that our generated views end up looking much like views that would have been specified directly in a GAV, LAV or GLAV framework.

The AutoMed intermediate query language IQL supports two kinds of operator for manipulating bags: the bag append operator, `++`, and also a family of operators which are all semantically based on a single function, `fold`.

`fold` applies a given function `f` to each element of a bag and then ‘folds’ a binary operator `op` into the resulting values. It is defined as follows, where `[]` is the empty bag, `[x]` is a singleton bag containing one element `x`, and `b1 ++ b2` is the union of two bags `b1` and `b2`:

```
fold f op e []           = e
fold f op e [x]         = f x
fold f op e (b1 ++ b2) = (fold f op e b1) op (fold f op e b2)
```

For example, `sum = fold (id) (+) 0` and `count = fold (lambda x.1) (+) 0`. The other common grouping and aggregation operators can also be defined in terms of `fold`, as can a function `flatmap`:

```
flatmap f b = fold f (++) [] b
```

`flatmap` can in turn be used to define selection, projection, and join operators. The comprehension syntax mentioned earlier also translates into successive applications of `flatmap`.

Optimisations for `fold` apply to all operators that can be defined in terms of it, including selection, projection, join, group-by, aggregation, and comprehensions. Regarding the view definitions generated from BAV pathways as described in the previous section, there are two particular optimisations that can be applied to them. Firstly, instances of `Void` can be removed. For the purposes of query processing, `Void` is regarded as being equal to the empty bag. Thus:

```
fold f op e Void = fold f op e []   = e
flatmap f Void   = flatmap f []     = []
group Void       = group []         = []
gc f Void        = gc f []          = []
sum Void         = sum []           = 0
```

```

count Void      = count []          = 0
Void ++ e       = [] ++ e = e ++ [] = e ++ Void = e
Void OR e       = e OR Void        = e

```

where `group` groups a bag of pairs according to common values in their first component and `gc f` groups a bag of pairs according to common values in their first component and then applies an aggregation function `f` to the resulting groups of second components. We refer the reader to [19] for more details of IQL and for references to work on fold-based functional query languages and optimisation techniques for such languages.

Due to the step-wise specification of our schema transformations, there is a second major optimisation which may be applicable. This is known as **loop fusion** and it replaces two successive iterations over a collection by one iteration provided the operators in question satisfy certain algebraic properties. A simple instance of loop fusion is the standard relational query optimisation  $\pi_A(\pi_B(R)) = \pi_{A,B}(R)$ . Loop fusion does not arise in the schema integration example of Section 2.1 but consider the following fragment of an AutoMed pathway. This first joins two schemes  $\langle\langle R, a \rangle\rangle$  and  $\langle\langle R, b \rangle\rangle$ , creating an intermediate relation  $\langle\langle I \rangle\rangle$ , and then projects onto the  $a$  and  $b$  attributes, creating a relation  $\langle\langle V \rangle\rangle$ , and finally deletes  $\langle\langle I \rangle\rangle$ :

```

addRel( $\langle\langle I \rangle\rangle$ , [(x, y, z) | (x, y) ←  $\langle\langle R, a \rangle\rangle$ ; (x, z) ←  $\langle\langle R, b \rangle\rangle$ ])
addRel( $\langle\langle V \rangle\rangle$ , [(y, z) | (x, y, z) ←  $\langle\langle I \rangle\rangle$ ])
deleteRel( $\langle\langle I \rangle\rangle$ , [(x, y, z) | (x, y) ←  $\langle\langle R, a \rangle\rangle$ ; (x, z) ←  $\langle\langle R, b \rangle\rangle$ ])

```

The view definition generated for  $\langle\langle V \rangle\rangle$  would be

$$[(y, z) | (x, y, z) ← [(x, y, z) | (x, y) ← \langle\langle R, a \rangle\rangle; (x, z) ← \langle\langle R, b \rangle\rangle]]$$

and the generator  $(x, y, z) ←$  in the outer comprehension can be fused with the head expression of the inner comprehension, giving:

$$[(y, z) | (x, y) ← \langle\langle R, a \rangle\rangle; (x, z) ← \langle\langle R, b \rangle\rangle]$$

There are a range of other standard algebraic optimisations that could be performed on the view definitions *e.g.* pushing down selections and projections. However, these kinds of optimisations will also be applied later, when a specific global query is reformulated by substituting into it the view definitions. Further optimisations and rewrites will be applied at this stage *e.g.* to bring constructs from the same local schemas together into sub-queries which can be posed entirely on one local schema and it is these sub-queries (appropriately translated) that will be sent to local data sources for evaluation.

## 5 Validating and Optimising Pathways

One important feature of the AutoMed approach is that once a set of schemas have been joined in a network of pathways, data and queries may be translated or migrated between any pair of schemas in the network. Such networks may be complex to analyse, so we need to support automated validation that a network is well-formed. We also need to support automated optimisation of the pathways between schemas, since they may contain redundant transformations.

To support such validation and optimisation of pathways, we have developed the **Transformation Manipulation Language (TML)** [23], which represents each transformation in a form suited to analysis of the schema constructs that are created, deleted or are required to be present for the transformation to be correct. Our definitions below require a function `sc` that given a query determines all schema constructs that must exist if the query is valid. For the IQL language constructs used in our earlier examples, `sc` is defined as:

$$\begin{aligned}
sc(\langle\langle r \rangle\rangle) &= \langle\langle r \rangle\rangle \\
sc(\langle\langle r, a \rangle\rangle) &= \{\langle\langle r \rangle\rangle, \langle\langle r, a \rangle\rangle\} \\
sc([q_1, \dots, q_n]) &= sc(q_1) \cup \dots \cup sc(q_n) \\
sc(q_1 ++ q_2) &= sc(q_1) \cup sc(q_2) \\
sc([q \mid q_1, \dots, q_n]) &= sc(q) \cup sc(q_1) \cup \dots \cup sc(q_n)
\end{aligned}$$

Note that as a shorthand, we will write the pair of queries  $q_l, q_u$  in **extend** or **contract** as just  $q$ , with the semantics in such cases that  $sc(q_l, q_u) = sc(q_l) \cup sc(q_u)$ . The TML formalises each transformation  $t_i$  of schema  $S_i$  into schema  $S_{i+1}$  as having four **conditions**  $a_i^+, b_i^-, c_i^+, d_i^-$ :

- The positive precondition  $a_i^+$  is the set of constructs that  $t_i$  implies must be present in  $S_i$ . It comprises those constructs that are present in the query of the transformation (given by  $sc(q)$ ) together with any constructs implied as being present by the construct  $c$ :

$$t_i \in \{\text{add}(c, q), \text{extend}(c, q)\} \rightarrow a_i^+ = (sc(c) - c) \cup sc(q)$$

$$t_i \in \{\text{delete}(c, q), \text{contract}(c, q), \text{rename}(c, c'), \text{id}(c, c')\} \rightarrow a_i^+ = sc(c) \cup sc(q)$$

- The negative precondition  $b_i^-$  is the set of constructs that  $t_i$  implies must not be present in  $S_i$ . It comprises those constructs which the transformation will add to the schema, and thus must not already be present:

$$t_i \in \{\text{add}(c, q), \text{extend}(c, q), \text{rename}(c', c), \text{id}(c', c)\} \rightarrow b_i^- = c$$

$$t_i \in \{\text{delete}(c, q), \text{contract}(c, q)\} \rightarrow b_i^- = \emptyset$$

- The positive postcondition  $c_i^+$  is the set of constructs that  $t_i$  implies must be present in  $S_{i+1}$ , and is derived in the same way as  $\overline{a_i^+}$  (i.e. the positive precondition of the  $\overline{t_i}$ ):

$$t_i \in \{\text{add}(c, q), \text{extend}(c, q), \text{rename}(c', c), \text{id}(c', c)\} \rightarrow c_i^+ = sc(c) \cup sc(q)$$

$$t_i \in \{\text{delete}(c, q), \text{contract}(c, q)\} \rightarrow c_i^+ = (sc(c) - c) \cup sc(q)$$

- The negative postcondition  $d_i^-$  is the set of constructs that  $t_i$  implies must not be present in  $S_{i+1}$ , and is derived in the same way as  $\overline{b_i^-}$ :

$$t_i \in \{\text{delete}(c, q), \text{contract}(c, q), \text{rename}(c, c'), \text{id}(c, c')\} \rightarrow d_i^- = c,$$

$$t_i \in \{\text{add}(c, q), \text{extend}(c, q)\} \rightarrow d_i^- = \emptyset$$

We can express  $LS_1 \rightarrow US_1$  in TML as shown below. Note that compound transformations such as  $t_1$  are first expanded into their primitive transformations before being converted into the TML.

$$\begin{aligned}
t_{1.1} &: [\emptyset, \{\langle\langle \text{student} \rangle\rangle\}, \{\langle\langle \text{student} \rangle\rangle\}, \emptyset] \\
t_{1.2} &: [\emptyset, \{\langle\langle \text{student, id} \rangle\rangle\}, \{\langle\langle \text{student} \rangle\rangle, \langle\langle \text{student, id} \rangle\rangle\}, \emptyset] \\
t_{1.3} &: [\emptyset, \{\langle\langle \text{student, id} \rangle\rangle\}, \{\langle\langle \text{student} \rangle\rangle, \langle\langle \text{student, sex} \rangle\rangle\}, \emptyset] \\
t_{1.4} &: [\emptyset, \{\langle\langle \text{student, id} \rangle\rangle\}, \{\langle\langle \text{student} \rangle\rangle, \langle\langle \text{student, dname} \rangle\rangle\}, \emptyset] \\
&\vdots \\
t_6 &: [\{\langle\langle \text{staff} \rangle\rangle, \langle\langle \text{staff, dname} \rangle\rangle\}, \{\langle\langle \text{dept} \rangle\rangle\}, \{\langle\langle \text{dept} \rangle\rangle, \langle\langle \text{staff} \rangle\rangle, \langle\langle \text{staff, dname} \rangle\rangle\}, \emptyset] \\
t_7 &: [\{\langle\langle \text{dept} \rangle\rangle\}, \{\langle\langle \text{dept, dname} \rangle\rangle\}, \{\langle\langle \text{dept} \rangle\rangle, \langle\langle \text{dept, dname} \rangle\rangle\}, \emptyset] \\
t_8 &: [\{\langle\langle \text{dept} \rangle\rangle\}, \{\langle\langle \text{dept, street} \rangle\rangle\}, \{\langle\langle \text{dept} \rangle\rangle, \langle\langle \text{dept, street} \rangle\rangle\}, \emptyset] \\
t_9 &: [\{\langle\langle \text{dept} \rangle\rangle\}, \{\langle\langle \text{dept, unname} \rangle\rangle\}, \{\langle\langle \text{dept} \rangle\rangle, \langle\langle \text{dept, unname} \rangle\rangle\}, \emptyset]
\end{aligned}$$

$t_{10} : [\{\langle\langle\text{staff}\rangle\rangle, \langle\langle\text{male}\rangle\rangle, \langle\langle\text{female}\rangle\rangle\}, \{\langle\langle\text{staff, sex}\rangle\rangle\}, \{\langle\langle\text{staff}\rangle\rangle, \langle\langle\text{staff, sex}\rangle\rangle, \langle\langle\text{male}\rangle\rangle, \langle\langle\text{female}\rangle\rangle\}, \emptyset]$   
 $t_{11} : [\{\langle\langle\text{male}\rangle\rangle, \langle\langle\text{male, id}\rangle\rangle\}, \emptyset, \{\langle\langle\text{male}\rangle\rangle\}, \{\langle\langle\text{male, id}\rangle\rangle\}]$   
 $t_{12} : [\{\langle\langle\text{male}\rangle\rangle, \langle\langle\text{staff}\rangle\rangle, \langle\langle\text{staff, sex}\rangle\rangle\}, \emptyset, \{\langle\langle\text{staff}\rangle\rangle, \langle\langle\text{staff, sex}\rangle\rangle\}, \{\langle\langle\text{male}\rangle\rangle\}]$   
 $t_{13} : [\{\langle\langle\text{female}\rangle\rangle, \langle\langle\text{female, id}\rangle\rangle\}, \emptyset, \{\langle\langle\text{female}\rangle\rangle\}, \{\langle\langle\text{female, id}\rangle\rangle\}]$   
 $t_{14} : [\{\langle\langle\text{female}\rangle\rangle, \langle\langle\text{staff}\rangle\rangle, \langle\langle\text{staff, sex}\rangle\rangle\}, \emptyset, \{\langle\langle\text{staff}\rangle\rangle, \langle\langle\text{staff, sex}\rangle\rangle\}, \{\langle\langle\text{female}\rangle\rangle\}]$

## 5.1 Well-formed Transformation Pathways

A pathway  $T$  from schema  $S_m$  to  $S_n$  is said to be **well-formed** if for each transformation step  $t_i : S_i \rightarrow S_{i+1}$  within it:

- The only difference between the schema constructs in  $S_{i+1}$  and  $S_i$  is those constructs specifically changed by transformation  $t_i$ , implying that  $S_{i+1} = (S_i \cup c_i^+) - d_i^-$  and  $S_i = (S_{i+1} \cup a_i^+) - b_i^-$
- The constructs required by  $t_i$  are in the schemas, implying that  $a_i^+ \subseteq S_i$ ,  $b_i^- \cap S_i = \emptyset$ ,  $c_i^+ \subseteq S_{i+1}$  and  $d_i^- \cap S_{i+1} = \emptyset$

The above definition leads to the recursive definition of a well-formed pathway,  $wf$ , given below. The first rule applies each transformation step in turn, and the second rule ensures that the schema that results from applying all the transformation steps is equal to the schema at the end of the pathway (equal both in terms of the schema constructs found in each schema and the extent of the schemas). Note that any implementation may use these rules in two ways. Firstly, given a schema  $S_m$  representing a data source, and pathway  $P$ , a new data source schema  $S_n$  and its extent can be derived. Secondly, if  $S_n$  exists as a data source already, a check can be made to verify that  $P$  correctly derives its schema and extent from that of  $S_m$ .

$$\begin{aligned}
 wf(S_m, S_n, [t_m, t_{m+1}, \dots, t_{n-1}]) &\leftarrow a_m^+ \subseteq S_m \wedge b_m^- \cap S_m = \emptyset \wedge \\
 &wf((S_m \cup c_m^+) - d_m^-, S_n, [t_{m+1}, \dots, t_{n-1}]) \\
 wf(S_m, S_n, []) &\leftarrow S_m = S_n \wedge Ext(S_m) = Ext(S_n)
 \end{aligned}$$

## 5.2 Reordering of Transformations

Certain transformations may be performed in any order, whilst others must be performed in a specific order. For example, in  $LS_1 \rightarrow US_1$ ,  $t_{11}$  must be performed before  $t_{12}$ , since the attribute  $\langle\langle\text{male, id}\rangle\rangle$  must be deleted before the  $\langle\langle\text{male}\rangle\rangle$  relation is deleted. However the sub-pathway  $t_{11}, t_{12}$  could be performed before or after the sub-pathway  $t_{13}, t_{14}$  since it does not matter which of the  $\langle\langle\text{male}\rangle\rangle$  or  $\langle\langle\text{female}\rangle\rangle$  relations is deleted first.

In the TML, this intuition is expressed by stating that transformations may be swapped provided the pathway remains well-formed. This may be verified by inspecting the conditions of each transformation. In particular, a pair of transformations  $t_i, t_{i+1}$  may be reordered to  $t_{i+1}, t_i$  provided:

1.  $t_i$  does not add a construct that is required by  $t_{i+1}$ , and  $\overline{t_{i+1}}$  does not add a construct that is required by  $\overline{t_i}$ , i.e.  $(c_n^+ - a_n^+) \cap a_{i+1}^+ = \emptyset$  and  $(a_{i+1}^+ - c_{i+1}^+) \cap c_n^+ = \emptyset$

2.  $t_i$  does not delete a construct required not to be present by  $t_{i+1}$ , and  $\overline{t_{i+1}}$  does not delete a construct required not to be present by  $\overline{t_i}$ , i.e.  $d_n^+ \cap b_{i+1}^+ = \emptyset$

We can now formalise the two examples given above from  $LS_1 \rightarrow US_1$ . For  $t_{11}, t_{12}$ , (1) is broken, and hence they may not be swapped. The changing of  $t_{11}, t_{12}, t_{13}, t_{14}$  to  $t_{13}, t_{14}, t_{11}, t_{12}$  may be performed by iteratively swapping pairs of transformations. Considering first  $t_{12}, t_{13}$ , we find neither rule is broken, and they may be reordered to  $t_{13}, t_{12}$ . Then  $t_{12}, t_{14}$  breaks neither rule, and may be reordered to  $t_{14}, t_{12}$ . This leaves a sub-pathway  $t_{11}, t_{13}, t_{14}, t_{12}$ , and a similar argument allows  $t_{11}$  swap with  $t_{13}$  and then  $t_{14}$ , to give the sub-pathway  $t_{13}, t_{14}, t_{11}, t_{12}$ .

### 5.3 Redundant Transformations

Two transformations  $t_x$  and  $t_y$  in a well-formed pathway  $T$  are **redundant** if  $T$  may be reordered such that  $t_x$  and  $t_y$  become consecutive within it, and  $T$  remains well-formed if they are then removed. Such redundant transformations will occur if a source schema evolves to model information in the same way as the global schema when previously it modelled the information in a different way. For example, suppose  $LS_1$  is evolved by transformations  $t_{49}, t_{50}, t_{51}, t_{52}, t_{53}$ , textually identical to transformations  $t_{10}, t_{11}, t_{12}, t_{13}, t_{14}$ , to model the gender of staff as a single `sex` attribute in a new version of the schema  $LS'_1$ . By reversing these transformation steps we can derive the pathway from the new to the old schema  $LS'_1 \rightarrow LS_1$ :

#### Example 5 Pathway $LS'_1 \rightarrow LS_1$

$\overline{t_{53}}$  addRel( $\langle\langle\text{female}\rangle\rangle, [x \mid (x, \text{'F'}) \leftarrow \langle\langle\text{staff, sex}\rangle\rangle]$ )  
 $\overline{t_{52}}$  addAtt( $\langle\langle\text{female, id}\rangle\rangle, [(x, x) \mid x \leftarrow \langle\langle\text{female}\rangle\rangle]$ )  
 $\overline{t_{51}}$  addRel( $\langle\langle\text{male}\rangle\rangle, [x \mid (x, \text{'M'}) \leftarrow \langle\langle\text{staff, sex}\rangle\rangle]$ )  
 $\overline{t_{50}}$  addAtt( $\langle\langle\text{male, id}\rangle\rangle, [(x, x) \mid x \leftarrow \langle\langle\text{male}\rangle\rangle]$ )  
 $\overline{t_{49}}$  deleteAtt( $\langle\langle\text{staff, sex}\rangle\rangle, [(x, \text{'M'}) \mid x \leftarrow \langle\langle\text{male}\rangle\rangle] ++ [(x, \text{'F'}) \mid x \leftarrow \langle\langle\text{female}\rangle\rangle]$ )

If we inspect the entire path  $LS'_1 \rightarrow US_1$ , consisting of  $LS'_1 \rightarrow LS_1$  followed by  $LS_1 \rightarrow US_1$ , it may be reordered to contain the sub-pathway:

$\overline{t_{51}}$  addRel( $\langle\langle\text{male}\rangle\rangle, [x \mid (x, \text{'M'}) \leftarrow \langle\langle\text{staff, sex}\rangle\rangle]$ )  
 $\overline{t_{50}}$  addAtt( $\langle\langle\text{male, id}\rangle\rangle, [(x, x) \mid x \leftarrow \langle\langle\text{male}\rangle\rangle]$ )  
 $\overline{t_{49}}$  deleteAtt( $\langle\langle\text{staff, sex}\rangle\rangle, [(x, \text{'M'}) \mid x \leftarrow \langle\langle\text{male}\rangle\rangle] ++ [(x, \text{'F'}) \mid x \leftarrow \langle\langle\text{female}\rangle\rangle]$ )  
 $t_{10}$  addAtt( $\langle\langle\text{staff, sex}\rangle\rangle, [(x, \text{'M'}) \mid x \leftarrow \langle\langle\text{male}\rangle\rangle] ++ [(x, \text{'F'}) \mid x \leftarrow \langle\langle\text{female}\rangle\rangle]$ )  
 $t_{11}$  deleteAtt( $\langle\langle\text{male, id}\rangle\rangle, [(x, x) \mid x \leftarrow \langle\langle\text{male}\rangle\rangle]$ )  
 $t_{12}$  deleteRel( $\langle\langle\text{male}\rangle\rangle, [x \mid (x, \text{'M'}) \leftarrow \langle\langle\text{staff, sex}\rangle\rangle]$ )

Clearly  $\overline{t_{49}}, t_{10}$  forms a redundant pair, because we are adding and deleting the same construct *with the same extent* since the query is the same. Once this has been performed  $\overline{t_{50}}, t_{11}$  may be removed for the same reason, and then  $\overline{t_{51}}, t_{12}$ . Once all other redundant pairs have been removed,  $LS'_1 \rightarrow US_1$  would comprise of just  $t_1$ – $t_9$ .

Using the TML, we can identify redundant transformations as satisfying:

$$(a_x^+ = c_y^+) \wedge (b_x^- = d_y^-) \wedge (c_x^+ = a_y^+) \wedge (d_x^- = b_y^-) \wedge Ext(c_x^+ \oplus a_x^+) = Ext(c_y^+ \oplus a_y^+)$$

where  $(x \oplus y) = (x - y) \cup (y - x)$ , and thus serves to find all the constructs being added or deleted by the pair of transformations. In practice, this rule means that any pair of transformations which add/extend and then delete/contract (in either order) the same construct are redundant, providing the query can be demonstrated to result in the same extent.

## 5.4 Partially Redundant Transformations

Two transformations  $t_x$  and  $t_y$  in a well-formed pathway  $T$  are **partially redundant** if  $T$  may be reordered to make  $t_x$  and  $t_y$  consecutive, and  $T$  remains well-formed if they are then replaced by a single transformation  $t_{xy}$ .

The pathway  $LS_1 \rightarrow LS_2$  has a pair of such partially redundant transformations, since it can be reordered to obtain the sub-pathway:

$$\begin{array}{l} t_7 \quad \text{addAtt}(\langle\langle \text{dept}, \text{dname} \rangle\rangle, [(x, x) \mid x \leftarrow \langle\langle \text{dept} \rangle\rangle]) \\ \overline{t_{18}} \quad \text{renameAtt}(\langle\langle \text{dept}, \text{dname} \rangle\rangle, \langle\langle \text{dept}, \text{deptname} \rangle\rangle) \end{array}$$

This may be replaced by the new transformation given below, which leaves a fully optimised pathway  $LS_1 \rightarrow LS_2$ .

$$t_{54} \quad \text{addAtt}(\langle\langle \text{dept}, \text{deptname} \rangle\rangle, [(x, x) \mid x \leftarrow \langle\langle \text{dept} \rangle\rangle])$$

Using the TML, we can identify partially redundant transformations as satisfying:

$$(a_x^+ = c_y^+ \oplus b_x^- = d_y^-) \wedge d_x^- \cap b_y^- = \emptyset \wedge d_x^- \neq \emptyset \wedge b_y^- \neq \emptyset$$

The simplifications for removing partially redundant and fully redundant transformations are summarised in the table below. The table shows what simplifications may be applied where a pair of transformations is found to operate on the same construct  $c$ . NWF denotes ‘not well-founded’ and  $[]$  denotes the removal of the pair. The table would remain correct if `extend` were to replace `add`, `contract` replace `delete`, and `id` replace `rename`. Further details of redundant and partially redundant transformations may be found in [23].

		$t_y$		
	$\text{add}(c, q)$	$\text{delete}(c, q)$	$\text{rename}(c, c')$	
$t_x$	$\text{add}(c, q)$	NWF	$[]$	$\text{add}(c', q)$
$\text{delete}(c, q)$	$[]$	NWF	$\text{delete}(c', q)$	NWF
$\text{rename}(c', c)$	NWF	$\text{delete}(c'', q)$	$[]$	$\text{rename}(c'', c')$
$\text{rename}(c'', c)$	NWF	$\text{delete}(c'', q)$	$\text{delete}(c'', q)$	$\text{rename}(c'', c')$

## 6 Concluding Remarks

In this paper we have described view generation and view optimisation in the AutoMed heterogeneous database integration framework. We have shown how the AutoMed schema pathways and views generated from them are amenable to considerable simplification, resulting in view definitions that look much like the views that would have been specified directly in a GAV, LAV or GLAV framework.

Since BAV integration is based on sequences of primitive schema transformations, it could be argued that data integration using it is more complex than with GAV, LAV or GLAV. However, the integration process can be greatly simplified by specifying well-known schema equivalences as higher-level composite transformations. We gave such an example, `extendTable`, in Section 2.1 above, and further examples are given in [17]. Moreover, we are working on techniques for semi-automatically generating transformation pathways to convert a source schema expressed in one modelling language into an equivalent target schema expressed in another modelling language, based on well known schema equivalences. We are also investigating schema matching techniques to automatically or semi-automatically integrate two specific schemas.

Finally, it should be noted that BAV is well-suited to peer-to-peer data integration (see [18]) since it lacks the directionality inherent in LAV, GAV and GLAV, all of which are tied to



the concept of there being a global schema which may not always be the case in peer-to-peer environments.

## References

- [1] M. Boyd, P.J. McBrien, and N. Tong. The automed schema integration repository. In *Proc. BNCOD02, LNCS 2405*, pages 42–45, 2002.
- [2] P. Buneman *et al.* Comprehension syntax. *ACM SIGMOD Record*, 23(1):87–96, 1994.
- [3] D. Calvanese, E. Damagio, G. De Giacomo, M. Lenzerini, and R. Rosati. Semantic data integration in P2P systems. In *Proceedings of DBISP2P*, Berlin, Germany, 2003.
- [4] S.S. Chawathe *et al.* The TSIMMIS project: Integration of heterogeneous information sources. In *Proc. 10th Meeting of the Information Processing Society of Japan*, pages 7–18, October 1994.
- [5] M. Friedman, A. Levy, and T. Millstein. Navigational plans for data integration. In *Proc. 16th National Conf. on AI*, pages 67–73. AAAI Press, 1999.
- [6] S. Kittivoravithkul. Transformation-based approach for integrating semi-structured data. Technical report, AutoMed Project, 2003.
- [7] M. Lenzerini. Data integration: A theoretical perspective. In *Proc. PODS02*, pages 247–258, 2002.
- [8] A.Y. Levy. Logic-based techniques in data integration. In J. Minker, editor, *Logic Based Artificial Intelligence*. Kluwer Academic Publishers, 2000.
- [9] A.Y. Levy, A.O. Mendelzon, Y. Sagiv, and D. Srivastava. Answering queries using views. In *Proc. PODS’95*, pages 95–104. ACM Press, May 1995.
- [10] A.Y. Levy, A. Rajamaran, and J. Ordille. Querying heterogeneous information sources using source description. In *Proc. VLDB’96*, pages 252–262, 1996.
- [11] J. Madhavan and A.Y. Halevy. Composing mappings among data sources. In *Proceedings of 29th Conference on VLDB*, pages 572–583, 2003.
- [12] I. Manolescu, D. Florescu, and D. Kossmann. Answering XML queries on heterogeneous data sources. In *Proc. VLDB’01*, pages 241–250, 2001.
- [13] P.J. McBrien and A. Poulouvasilis. Automatic migration and wrapping of database applications — a schema transformation approach. In *Proc. ER’99, LNCS 1728*, pages 96–113, 1999.
- [14] P.J. McBrien and A. Poulouvasilis. A uniform approach to inter-model transformations. In *Proc. CAiSE’99, LNCS 1626*, pages 333–348, 1999.
- [15] P.J. McBrien and A. Poulouvasilis. A semantic approach to integrating XML and structured data sources. In *Proc. CAiSE’01, LNCS 2068*, pages 330–345, 2001.

- [16] P.J. McBrien and A. Poulouvasilis. Schema evolution in heterogeneous database architectures, a schema transformation approach. In *Proc. CAiSE'02, LNCS 2348*, pages 484–499, 2002.
- [17] P.J. McBrien and A. Poulouvasilis. Data integration by bi-directional schema transformation rules. In *Proc. ICDE'03*, 2003.
- [18] P.J. McBrien and A. Poulouvasilis. Defining peer-to-peer data integration using both as view rules. In *Proceedings of DBISP2P*, Berlin, Germany, 2003.
- [19] A. Poulouvasilis. The AutoMed Intermediate Query Language. Technical report, AutoMed Project, 2001.
- [20] A. Poulouvasilis and P.J. McBrien. A general formal framework for schema transformation. *Data and Knowledge Engineering*, 28(1):47–71, 1998.
- [21] M.T. Roth and P. Schwarz. Don't scrap it, wrap it! A wrapper architecture for data sources. In *Proc. VLDB'97*, pages 266–275, Athens, Greece, 1997.
- [22] M. Templeton, H.Henley, E.Maros, and D.J. Van Buer. InterViso: Dealing with the complexity of federated database access. *The VLDB Journal*, 4(2):287–317, 1995.
- [23] N. Tong. Database schema transformation optimisation techniques for the AutoMed system. Technical report, AutoMed Project, 2002.
- [24] D. Williams. Representing RDF and RDF Schema in the HDM. Technical report, Automed Project, 2002.