# Effective Data Integration in the Presence of Temporal Semantic Conflicts

Hongwei Zhu
Stuart E. Madnick
Michael D. Siegel

MIT Sloan School of Management
30 Wadsworth Street, Cambridge, MA 02142, USA
{mrzhu, smadnick, msiegel}@mit.edu

Working Paper CISL# 2004-02

February 2004

Composite Information Systems Laboratory (CISL)
Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02142

This page is blank

# Effective Data Integration in the Presence of Temporal Semantic Conflicts

Hongwei Zhu, Stuart E. Madnick, Michael D. Siegel
MIT Sloan School of Management
30 Wadsworth Street, Cambridge, MA 02142
{mrzhu, smadnick, msiegel}@mit.edu

**Abstract**

*The change in meaning of data over time poses significant challenges for the use of that data. These challenges exist in the use of an individual data source and are further compounded with the integration of multiple sources. In this paper, we identify three types of temporal semantic heterogeneities, which have not been addressed by existing research. We propose a solution that is based on extensions to the Context Interchange framework. This approach provides mechanisms for capturing semantics using ontology and temporal context. It also provides a mediation service that automatically resolves semantic conflicts. We show the feasibility of this approach by demonstrating a prototype that implements a subset of the proposed extensions.*

## 1 Introduction

Effective management of temporal data has become increasingly important in application domains from day-to-day record keeping to counter-terrorism efforts. In some cases, it is even required by law for organizations to store historical data and make sure it is accurate and easy to retrieve[1]. While temporal databases can be used to manage the data, ensuring that the retrieved data is meaningful to the users is still an unsolved problem when data semantics changes over time.

As an example, suppose an arbitrage analyst in New York needs to compare Daimler-Chrysler's stock prices at New York and Frankfurt exchanges. He retrieved the data from Yahoo's online historical database, see Figure 1. Two strange things caught his eyes at a quick glance at the data. First, prices at two exchanges differ substantially; and second, the price at Frankfurt dropped by almost 50% at the turn from 1998 to 1999! Imagine what conclusions the analyst would draw from this data.

| Date | Open | High | Low | Close | Volume | Adj Close* |
|---|---|---|---|---|---|---|
| 8-Jan-99 | 105.25 | 105.62 | 103.86 | 104.28 | 1,112,600 | 90.55 |
| 7-Jan-99 | 102.21 | 105.25 | 102.21 | 104.89 | 1,505,400 | 91.08 |
| 6-Jan-99 | 105.25 | 105.74 | 103.92 | 105.13 | 2,061,200 | 91.29 |
| 5-Jan-99 | 99.05 | 103.43 | 98.93 | 103.31 | 2,634,600 | 89.70 |
| 4-Jan-99 | 99.66 | 100.69 | 98.08 | 98.99 | 3,441,400 | 85.95 |
| 31-Dec-98 | 94.49 | 94.55 | 93.21 | 93.51 | 506,900 | 81.20 |
| 30-Dec-98 | 94.97 | 95.34 | 94.18 | 94.18 | 391,300 | 81.78 |
| 29-Dec-98 | 96.13 | 96.25 | 95.64 | 95.95 | 1,195,700 | 83.31 |
| 28-Dec-98 | 95.64 | 96.43 | 95.16 | 95.64 | 1,707,800 | 83.05 |
| 24-Dec-98 | 91.38 | 91.87 | 90.47 | 91.69 | 380,400 | 79.61 |
| 23-Dec-98 | 92.05 | 92.17 | 90.59 | 91.51 | 1,352,900 | 79.45 |
| 22-Dec-98 | 90.11 | 90.29 | 89.13 | 90.04 | 1,412,300 | 78.19 |
| 21-Dec-98 | 90.04 | 90.35 | 89.25 | 89.56 | 1,831,700 | 77.76 |

* Close price adjusted for dividends and splits.

| Date | Open | High | Low | Close | Volume | Adj Close* |
|---|---|---|---|---|---|---|
| 8-Jan-99 | 91.90 | 93.70 | 91.20 | 92.50 | 677,385 | 90.71 |
| 7-Jan-99 | 93.40 | 93.40 | 89.50 | 90.70 | 978,738 | 88.94 |
| 6-Jan-99 | 90.10 | 92.40 | 89.30 | 92.30 | 13,950,500 | 90.51 |
| 5-Jan-99 | 86.80 | 88.60 | 86.10 | 86.80 | 12,329,300 | 85.12 |
| 4-Jan-99 | 83.50 | 88.50 | 82.50 | 87.50 | 13,660,200 | 85.81 |
| 30-Dec-98 | 166.30 | 167.90 | 164.50 | 164.50 | 4,934,820 | 161.32 |
| 29-Dec-98 | 166.00 | 166.50 | 164.50 | 165.00 | 5,039,660 | 161.81 |
| 28-Dec-98 | 159.50 | 167.80 | 159.30 | 166.50 | 9,748,480 | 163.28 |
| 23-Dec-98 | 154.30 | 159.90 | 154.30 | 157.70 | 11,600,100 | 154.65 |
| 22-Dec-98 | 154.30 | 156.20 | 153.70 | 154.00 | 10,870,600 | 151.02 |
| 21-Dec-98 | 152.00 | 155.10 | 151.30 | 154.30 | 14,728,200 | 151.31 |

* Close price adjusted for dividends and splits.

Figure 1. Historical stock prices for Daimler-Chrysler. *Top*: New York; *Bottom*: Frankfurt

These unusual observations result from unresolved semantic conflicts between the data sources and the data receiver. In this case, not only are the currencies for stock prices different at the two exchanges, but the currency at Frankfurt also changed from German Marks to Euros at the beginning of 1999. Once the data is normalized using this knowledge, it can be seen there is neither significant arbitraging opportunity nor abrupt price plunge for this stock. We call metadata knowledge such as the currency for price *context knowledge*, and the history of time varying metadata *temporal context*.

[1] See Robert Sheier on "Regulated storage" in Computer World, **37**(46), November 17, 2003. *Health Insurance Portability Act* requires healthcare providers keep records till two years after death of patients; *Sarbanes-Oxley Act* requires auditing firms retain records of financial statements.

To allow data receivers like the analyst to effectively use data from temporally heterogeneous sources, we need to represent temporal context knowledge and incorporate it into data integration and query answering systems. This is a challenging problem not addressed by previous research. Temporal database research has focused on management of temporal data in a homogeneous environment, providing no mechanism for semantic annotation. Semantic data integration techniques developed so far are based on snapshot data models that ignore the time dimension.

The objective of this research is to fill this gap by developing semantic integration techniques to effectively resolve temporal semantic conflicts between data sources and receivers. Specifically, we extend the Context Interchange (COIN) framework [6, 10] with temporal contextual knowledge representation and reasoning capability to allow for effective retrieval of data from heterogeneous sources.

We begin in the next section with an example to illustrate various temporal semantic heterogeneities. In Section 3 we give a brief review of related research. We outline our approach to these challenges in section 4 and present some preliminary results in Section 5. In the final section, we summarize and briefly discuss future research.

## 2 Challenges of temporal data integration

### 2.1 A simple integration example

A temporal database is one that supports some aspect of time, not counting user-defined time such as birthday and hiring date [13]. This rather informal definition is due to the fact that the temporal dimensions are often application specific, therefore it is either difficult or unnecessary to support all aspects of time. Nevertheless, most *temporal data* can be viewed as time-stamped propositions and represented as relational tuples with timestamps. Table 1 gives an example of some time series data for a company.

Intuitively, the example describes how the values of several attributes change over time. Each tuple represents a fact that can be viewed as a predicate with a timestamp argument and other non-temporal arguments. However, there are other unspecified metadata attributes, such as currency type and scale factor, that critically determine the truth value of each predicate. We call the specification of metadata attributes *context* knowledge. For metadata attributes whose value changes over time, a specification of their history is termed *temporal context*.

Table 2 gives examples of the context knowledge in a simple integration scenario involving the source in Table 1 and a possible receiver. The receiver context can be time varying as well. Semantic conflicts arise because the source and the receiver have different contexts, which need to be reconciled for the receiver to meaningfully use the data. Imagine the complexity of integration scenarios that involve tens or even hundreds of sources and receivers with time varying heterogeneous contexts. We need integration technologies that can effectively manage this complexity.

Table 1. Company time series data

| Year | Num_Employee | Profit | Tax |
|---|---|---|---|
| … | | | |
| 1999 | 5100 | 4.2 | 1.1 |
| 2000 | 12000 | 13000 | 2500 |
| 2001 | 25.3 | 20000 | 4800 |
| 2002 | 30.6 | 35.3 | 7.97 |
| … | | | |

Table 2. Examples of temporal context knowledge

| | Source | Receiver |
|---|---|---|
| Currency | Francs(FRF), year ≤ 2000<br>Euros, year≥2001 | USD, always |
| Scale factor for **profit** and **tax** | 1M , year =< 1999<br>1K, 2000≤year≤2001<br>1M, year≥2002 | 1K, always |
| Scale factor for **Num_Employee** | 1, year≤2001<br>1K, year≥2002 | 1K, always |
| **Profit** | Exclude tax, year≤2000<br>Include tax, year≥2001 | Include tax, always |

### 2.2 Temporal semantic heterogeneities

We see at least three categories of issues in the integration of temporal data. In the following, we informally describe each using examples.

*Representational heterogeneity* – the same relational attribute may be represented differently in the time span of a data source. In addition to *currency* changes for monetary concepts like *profit* and *tax*, there are also *scale factor* changes, as described in Table 2.

*Ontological heterogeneity* – the ontological concept represented by an attribute may change over time. There are many examples of this kind. In Table 2, profit on and before 2000 excludes taxes, afterwards it includes taxes. There are also cases where the entity referred to by an identifier changes over time. For example, stock symbol

"C" at New York Stock Exchange (NYSC) used to refer to Chrysler but changed to refer to Citigroup on December 4, 1998 after Chrysler merged with Daimler-Benz and the merged company chose to use symbol "DCX". Similarly, country code "YUG" for Yugoslavia have different geographic boundaries at different times of the Balkans war.

The derivation method or composition of complex concepts often changes over time. Many government published time series data sets often come with a "Change of Definitions" that explains changes to terminologies. For example, the national unemployment data may include undocumented immigrants in the workforce at one time and exclude them at another.

*Heterogeneity in temporal entity* – the abstraction and representation of time domain differs across systems and time. Although a temporal entity is just another data type, it has special properties and operations that warrant a category of its own. The example in Table 1 uses point representation for the timestamp attribute *year*. Another system may choose to use intervals, e.g., [1/1/1999, 12/31/1999] for the year 1999. Differences in calendar systems, time zones, and granularities present many challenges for integration.

The semantics of the association between propositions described by the non-temporal attributes in a tuple and the temporal entity may differ across attributes. How the truth of a proposition over an interval is related to its truth over subintervals is described by the proposition's *heredity* properties [23]. Recognizing this property is useful for temporal query language design. For example, if in a bank account database the balance over an interval is known and the user queries the balance at a time within the interval, the query language should use the *liquidity* property of balance attribute to infer the result [3]. We observe that heredity is often attribute dependent and does not change over time or across data sources. Thus we need not consider heterogeneity of this property in the data integration setting.

In an effective integration framework, data receivers should not be burdened by these context heterogeneities; rather, it should be a service of the system to record and use context knowledge to reconcile context differences before delivering data to the receiver. Our temporal extension to COIN framework will provide such a solution.

## 3 Review of Related Research

Related research can be found in the areas of temporal database, temporal reasoning, and data integration. Although it provides useful concepts and techniques, research from each area alone does not address all temporal semantic heterogeneity problems identified in this paper. The following brief review is not intended to summarize or criticize the findings in each area; rather, it is to identify the most relevant results and show what is missing from a temporal semantic data integration point of view.

### 3.1 Temporal databases

In temporal database research, the time domain is often represented as time points with certain granularities. An interval is a set of contiguous points and can be represented as a pair of begin and end points. A time point represents a segment on the continuous time line, thus it has a duration and is not an instant in time ontology [11]. Although both are indivisible, *point* type has a successor function while *instant* type does not.

Over 40 temporal data models have been proposed [20], yet only a few have been implemented with very limited commercial impact. Many of the models let the system manage timestamps, which effectively hide the timestamp attribute from the user. This approach is inconsistent with relational theory, as articulated in [4], where they explicitly make the timestamp an attribute of the relational schema. As commonly practiced, databases that store temporal data often have a schema with explicit timestamp attribute(s); standard SQL is used to retrieve data and temporal operations are selectively implemented in the application layer. Our framework will target the common situation where data sources have limited temporal support.

As in the case of conventional databases, temporal databases also fail to facilitate context knowledge management. As a result, context is often hard-coded into data transformations in multi-dimensional data warehouses. This ad-hoc approach lacks flexibility and scalability.

### 3.2 Temporal reasoning

While a restricted set of temporal logics can be executed using logic programming, there

seems to be a trend where temporal logics are transformed into temporal constraints to take advantage of the efficiency of constraint solvers. The framework provided in [17] combines qualitative and quantitative (metric) temporal relations over both time points and time intervals. These relationships can be considered as temporal constraints in constraint logic programming. Therefore, temporal reasoning can be treated as a constraint solving problem, to which a number of constraint solving techniques [12] can be applied. A solver implemented using constraint handling rules (CHR) [9] is demonstrated in [7]. We will exploit this general approach in this research.

Temporal granularity research, both logic [18] and algebraic [3] based, has developed techniques for representing and reasoning about granularities and user-defined calendars. Conversions between granularities [2] will be useful in dealing with heterogeneities in temporal entity.

### 3.3 Data integration

Approaches to achieving data integration largely fall into tight-coupling and loose-coupling categories depending on whether a global schema is used [6, 10]. In these approaches, intended data semantics in sources and receivers are explicitly incorporated in either the view definitions or the user queries. The computation complexity in rewriting user queries for the former approach [15] and the user's burden of writing complex queries for the latter significantly limit the flexibility and scalability of these approaches.

A middle ground approach, as articulated in [5] and demonstrated in [6, 10], avoids these shortcomings by encapsulating data semantics into a context theory and maintaining accessibility of source schema by users. In the COIN approach, the user issues queries as if all sources are in the user's context and a mediator is used to rewrite the queries to resolve semantic differences.

Unfortunately, existing approaches, including COIN, assume static metadata and ignore the temporal aspect of context knowledge. Consequently, temporal concepts are missing in the ontologies used in these systems. Research in medical information systems concerns what diagnostic conclusions can be drawn from a set of temporal data (*e.g.*, a series of high temperatures lead to "having a fever" conclusion). In a seemingly relevant research [19, 22], interpretation contexts are dynamically constructed using temporal data as input (*e.g.*, from doses of a certain drug infer a certain treatment protocol, which has a side effect of causing high temperatures. In this *interpretation context*, high temperatures do not imply a fever). However, it also assumes that interpretation contexts do not vary in time.

This research will focus on the representation and reasoning of temporal context. We will develop a framework that incorporates context knowledge into query evaluation process to automatically detect and reconcile temporal semantic conflicts. By combining the concepts and techniques from the three relevant research areas, we develop a scalable solution to temporal heterogeneity problems.

## 4. Temporal COIN approach

A recent extension to the COIN framework demonstrates its new capability in solving ontological heterogeneity problems [6]. With various temporal representation and reasoning techniques at our disposal, we feel that COIN can be further extended to handle temporal semantic heterogeneities.

### 4.1 The COIN framework

The COIN framework [6] uses an *ontology* as a formal mechanism for representing context knowledge and a *mediation* service to dynamically detect and reconcile semantic conflicts in user queries. Implemented in abductive constraint logic programming (ACLP) [14], the mediator generates *mediated queries* (*MQs*) that serve as intensional answers to the user. A distributed query executioner generates a query plan and brings extensional answers to the user. *Conversion functions* defined internally or externally are called during query execution stage to convert extensional dataset from its source context into receiver's context.

The existing COIN uses a snapshot data model that does not allow temporal context representation; the mediator also lacks temporal reasoning capability.

### 4.2 Temporal extensions to representation

The extended framework admits temporal data sources, which are assumed to be relational with an explicit timestamp attribute in their schema. They accept SQL queries with usual

comparison operators (=, >, <, *etc.*) on timestamp domain.

**Definition** An *ontology* defines a common semantic domain that consists of semantic data types and their relationships, including 1) *is-a* relation, indicating a type is a subtype of another; and 2) named attribute relation, indicating the domain and range types of the attribute.

The ontology is augmented with temporal concepts as defined in the time ontology from an early DAML effort [11]. The most general one is called *Temporal Entity*, which can be further specialized as *Instant* or *Interval*. Each element in the source schema is mapped to a corresponding semantic data type in the ontology by an *elevation axiom*. A timestamp can be elevated to Temporal Entity or a finer type in the ontology. For types whose values are time dependent, we relate them to a temporal entity type via *temporal attribute*.

There are two kinds of attributes in the ontology. The range of a regular attribute usually is defined in source schema and obtains its value from the extensional database (EDB). A *contextual attribute*, also called a *modifier*, is an implicit attribute whose value functionally determines the interpretation and the value of the regular attribute [21]. For example, **currency** and **scale factor** in Table 1 are contextual attributes for **profit** and **tax**.

**Definition** The *temporal context* of a data source or a receiver is a specification of the history of all contextual attributes in the ontology. Mathematically, it is set of <*contextual_attribute, history*> pairs, where *history* is a set of <*value, valid_interval*> pairs.

In existing COIN, a context is simply a set of <*contextual_attribute, value*> pairs. The temporal extension allows us to represent the entire history of each contextual attribute. If the value does not change over time, the *history* set is simply a singleton with the *valid_interval* covering entire time span of interest. Our implementation achieves backward compatibility by treating <*contextual_attribute, value*> as the shorthand for <*contextual_attribute, {<value, entire_time_span>}*>.

Temporal entity type may also have contextual attributes, *e.g.*, granularity, time zone, etc., to account for various contexts.

### 4.3 Temporal extensions to mediation

Given a user query expressed as if everything were in user's context, the mediator detects and reconciles semantic differences by examining the assignments to all contextual attributes and applying appropriate conversion functions if the assignment differs between any source and the receiver. With temporal extensions, contextual attributes are no longer singly valued. However, at any point in time, there is only one valid value for each attribute. The mediator needs to find the maximum time interval over which all involved contextual attributes are singly valued. Over this interval, a mediated query can be generated as in the case of existing COIN; the derived interval appears in the $\mathcal{MQ}$ as additional constraints over the attribute of temporal entity type.

The mediator translates *history* pairs for contextual attributes into temporal constraints, which are posted into a constraint store concurrently solved by solvers written in CHR. Through back tracking, all intervals over which contextual attributes are singly valued are found.

In our framework, contexts are declaratively defined using First Order Logic rules. The mediator is a general abductive reasoner. When new sources are added, we only need to add context rules for them. External functions can also be called to convert between contexts. These features lend COIN great flexibility and scalability.

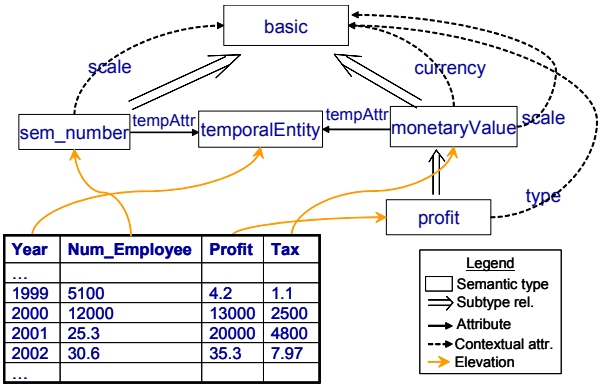
## 5 Prototype and preliminary results

We are able to solve a range of temporal heterogeneity problems exemplified in Section 2 by implementing a fraction of the suggested extensions.

### 5.1 Representation

Figure 2 shows the ontology we constructed for the example. Here, we use the most general concept *temporal entity*, which can represent both instants and intervals.

Using elevation axioms, we create the mappings between attributes in each data source and the types in the ontology, as shown in Figure 2.




| Year | Num_Employee | Profit | Tax |
|---|---|---|---|
| ... | | | |
| 1999 | 5100 | 4.2 | 1.1 |
| 2000 | 12000 | 13000 | 2500 |
| 2001 | 25.3 | 20000 | 4800 |
| 2002 | 30.6 | 35.3 | 7.97 |
| ... | | | |

Figure 2. Example ontology and elevation

We model the time line as discrete and unbounded with both points and intervals as primitives. The past and future infinities are represented by constants *bottom* and *top*. Intervals are represented with begin and end points, which can be accessed by functions *begins* and *ends*. We implement the ≤ relation between points as a *tle* constraint. The *contains* relation between an interval and a point is translated into *tle* constraints; the *overlaps* relation between intervals are also translated into *tle* constraints.

This simple model has all the expressive power to represent the kinds of temporal knowledge in Table 2. For example, internally we use the following Prolog statements to represent the source context for currency:

```
modifier(monetaryValue,O,currency,c_src,M):-
     containObj([bottom, 2000], O),
     cste(basic, M, c_src, "FRF").
modifier(monetaryValue,O,currency,c_src,M):-
     containObj([2001, top],O),
     cste(basic, M, c_src, "EUR").
```

The head of the statement reads: $O$ of type *monetaryValue* has a contextual attribute (*i.e.*, modifier) *currency*, whose value in source context *c_src* is $\mathcal{M}$. Its body has two predicates. *containObject(I, O)* will use the *tempAttr* of $O$ to obtain its temporal attribute $\mathcal{T}$ (which corresponds to *Year* attribute in the EDB) of type *temporalEntity* and add constraint *contains(I, T)*. The helper predicate *cste* specifies the primitive value of $\mathcal{M}$ in *c_src* context. Thus, the history of each contextual attribute is now a set of pairs $<\mathcal{V}_i, I_i>$, where $\bigcup_i I_i = [bottom, top]$.

For context knowledge that does not change over time, we could have used *[bottom, top]* interval in *containObj* predicate. Since the translated constraints are always true, we will not include this predicate for this case.

### 5.2 Mediation

As described earlier, the mediation service needs to find the maximum interval over which all contextual attributes are singly valued. Figure 3 helps visualize this task by graphically representing the context knowledge in Table 2. For example, *[bottom, 1999]* is such an interval where the source context can be described with a set of singly valued contextual attributes:

```
{<monetaryValue.currency, "FRF">,
 <monetaryValue.scale, "1000000">,
 <profit.type, "taxExluded">,
 <sem_num.scale, "1">}.
```

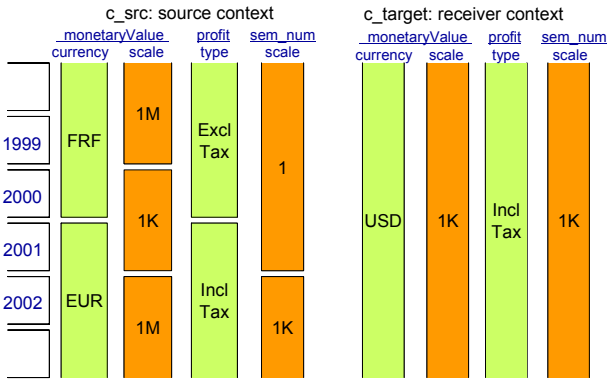



Figure 3. Visualization of temporal context

Recall that we translate all temporal relations into *tle* constraint over points. Each contextual attribute will generate two *tle* constraints for the temporal variable. The above problem is thus turned into to a problem of solving the constraints generated by all the contextual attributes, which is solved concurrently using a solver implemented in CHR.

Constraints over *bottom* and *top* can be removed using simplification rules so that these two literals do not appear in the list of abducted predicates. Constraints over other time points can be pair-wise simplified. We also implement *overlaps* to simplify *tle* constraints over four points at a time. These rules tighten the bounds of the temporal variable or signify a failure if inconsistencies are found.

Along with rules that handle equality constraint, this point based temporal constraint solver covers the 13 relations for temporal intervals in Allen [1]. Relations *before*, *after*, *meets*, and *met_by*

generate a failure, all the rest relations are subsumed into *overlaps*.

Through backtracking, the recursive algorithm finds all intervals over which contextual attributes are singly valued. Conversions are applied as in the case of existing COIN implementation. This simple temporal constraint based extension transforms a temporal context problem into a set of non-temporal problems, thereby allows us to reuse the non-temporal implementation of the COIN mediator.

### 5.3 Preliminary results.

These temporal extensions to COIN framework allow us to achieve semantic interoperability for the integration example. The prototype can generate *MQs* that reconcile temporal context differences.

As an example, suppose a user in the receiver context wants to retrieve data from the company time series relation named *Financials* in Table 1. So he issues the following SQL query:

```
Select Year, Num_Employee, Profit
From Financials;
```

and expects the returned data to be in his context. The query is translated into a well formed Datalog query in our prototype system. The extended COIN mediator takes this query and the representation of the integration as input, and produces the following mediated query in Datalog (which COIN eventually converts to SQL):

```
%1. =<1999: currency,scale,type;scale
answer('V32', 'V31', 'V30') :-
        'V29' is 'V28' * 1000.0,
        'V31' is 'V27' * 0.001,
        olsen("FRF", "USD", 'V26', 'V32'),
        'V28' is 'V25' * 'V26',
        financials('V32', 'V27', 'V25', 'V24'),
        'V32' =< 1999,
        'V23' is 'V24' * 'V26',
        'V22' is 'V23' * 1000.0,
        'V30' is 'V29' + 'V22'.
%2. 2000: currency and type;scale
answer(2000, 'V21', 'V20') :-
        'V21' is 'V19' * 0.001,
        financials(2000, 'V19', 'V18', 'V17'),
        olsen("FRF", "USD", 'V16', 2000),
        'V15' is 'V18' * 'V16',
        'V14' is 'V17' * 'V16',
        'V20' is 'V15' + 'V14'.
%3. 2001: currency;scale
answer(2001, 'V13', 'V12') :-
        'V13' is 'V11' * 0.001,
        financials(2001, 'V11', 'V10', 'V9'),
        olsen("EUR", "USD", 'V8', 2001),
        'V12' is 'V10' * 'V8'.
%4. >=2002: currency,scale;none
answer('V7', 'V6', 'V5') :-
        olsen("EUR", "USD", 'V4', 'V7'),
        financials('V7', 'V6', 'V3', 'V2'),
        2002 =< 'V7',
        'V1' is 'V3' * 'V4',
        'V5' is 'V1' * 1000.0.
```

The mediated query has four subqueries, each resolves a set of semantic conflicts that exist in the time specified by the timestamp attribute. Note that *olsen* predicate corresponds to a currency conversion data source introduced by the conversion function for *currency* contextual attribute. Readers are encouraged to walkthrough the subqueries to verify that they resolve all the semantic conflicts in Table 2 or in Figure 3.

## 6. Discussion and future plan

We have identified three types of semantic heterogeneity in the integration of temporal data. There is an ever increasing need to efficiently handle these heterogeneities as more historical data is used for auditing, mining, forecasting, investigation, and many other purposes.

We proposed temporal extensions to the COIN framework. A prototype of the extensions shows that our approach is capable of solving temporal context problems. With its declarative knowledge rule base and its capability of calling external functions, this approach is flexible and extensible.

Our future research aims to develop this approach in several aspects. Current representation of temporal context explicitly compares an interval with the temporal attribute of an object. The representation may be made cleaner by using an annotated temporal constraint logic [8]. We need to investigate how this logic can be integrated with the ACLP implementation of COIN mediator.

The most important part of future research will be focused on the heterogeneities of temporal entities. Intuitively, we can imagine a solution that adds various contextual attributes to the temporal entity type in the ontology and relies on external functions to convert between contexts. If this is not expressive enough to represent the diversity of time, a richer time ontology may be necessary. We also need to incorporate metric temporal reasoning, which often involves computations of one or more calendars. We will investigate the feasibility of leveraging web services like those in [2]. This is a challenging and important research area because misunderstanding date and time can have serious consequences, as history has shown in an 1805 event [16] where the Austrian troops were

forced to surrender largely because of the misunderstanding of a date in two different calendar systems.

**Acknowledgements**: The study has been supported, in part, by MITRE Corp., MIT-MUST project, the Singapore-MIT Alliance, and Suruga Bank.

## References

[1] J. F. Allen, "Maintaining Knowledge about Temporal Intervals," *Communications of the ACM*, vol. 26, pp. 832-843, 1983.

[2] C. Bettini, "Web services for time granularity reasoning," presented at TIME-ICTL'03, 2003.

[3] C. Bettini, S. Jajodia, and X. S. Wang, *Time Granularities in Databases, Data Mining, and Temporal Reasoning*: Springer, 2000.

[4] C. J. Date, H. Darwen, and N. A. Lorentzos, *Temporal Data and the Relational Model: A Detailed Investigation into the Application of Interval Relation Theory to the Problem of Temporal Database Management*: Morgan Kaufmann Publishers, 2003.

[5] A. Farquhar, A. Dappert , R. Fikes, and W. Pratt, "Integrating Information Sources Using Context Logic," presented at AAAI Spring Symposium on Information Gathering from Distributed Heterogeneous Environments, 1995.

[6] A. Firat, "Information Integration using Contextual Knowledge and Ontology Merging," PhD Thesis, MIT, 2003.

[7] T. Frühwirth, "Temporal Reasoning with Constraint Handling Rules," ECRC-94-5, 1994.

[8] T. Frühwirth, "Temporal Annotated Constraint Logic Programming," *Journal of Symbolic Computation*, vol. 22, pp. 555-583, 1996.

[9] T. Frühwirth, "Theory and Practice of Constraint Handling Rules," *Journal of Logic Programming*, vol. 37, pp. 95-138, 1998.

[10] C. H. Goh, S. Bressan, S. Madnick, and M. Siegel, "Context Interchange: New Features and Formalisms for the Intelligent Integration of Information," *ACM TOIS*, vol. 17, pp. 270-293, 1999.

[11] J. R. Hobbs, "A DAML Ontology of Time," 2002.

[12] J. Jaffar and M. J. Maher, "Constraint Logic Programming: A Survey," *Journal of Logic Programming*, vol. 19/20, pp. 503-581, 1999.

[13] C. S. Jensen and e. al., "The Consensus Glossary of Temporal Database Concepts---February 1998 Version," 1998.

[14] A. C. Kakas, A. Michael, and C. Mourlas, "ACLP: Integrating Abduction and Constraint Solving," *Journal of Logic Programming*, vol. 44, pp. 129-177, 2000.

[15] M. Lenzerini, "Data integration: a theoretical perspective," presented at 21st ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, 2002.

[16] S. E. Madnick, "Metadata Jones and the Tower of Babel: The Challenge of Large-Scale Semantic Heterogeneity," presented at 1999 IEEE Meta-Data Conference, 1999.

[17] I. Meiri, "Combining Qualitative and Quantitative Constraints in Temporal Reasoning," presented at AAAI 91, 1996.

[18] A. Montanari, "Metric and Layered Temporal Logic for Time Granularity," PhD Thesis, University of Amsterdam, 1996.

[19] J. H. Nguyen, Y. Shahar, S. W. Tu, A. K. Das, and M. A. Musen, "Integration of Temporal Reasoning and Temporal-Data Maintenance Into A Reusable Database Mediator to Answer Abstract, Time-Oriented Queries: The Tzolkin System," *Journal of Intelligent Information Systems*, vol. 13, pp. 121-145, 1999.

[20] G. Özsoyoglu and R. T. Snodgrass, "Temporal and Real-Time Databases: A Survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 7, pp. 513-532, 1995.

[21] E. Sciore, M. Siegel, and A. Rosenthal, "Using Semantic Values to Facilitate Interoperability Among Heterogeneous Information Systems," *ACM TODS*, vol. 19, pp. 254-290, 1994.

[22] Y. Shahar, "A Knowledge-Based Method for TemporalL Abstraction of Clinical Data," Stanford University, 1994.

[23] Y. Shoham, "Temporal Logics in AI: Semantical and Ontological Considerations," *Artificial Intelligence*, vol. 33, pp. 89-104, 1987.