Effective Data Integration in the Presence of Temporal Semantic Conflicts

Hongwei Zhu, Stuart E. Madnick, Michael D. Siegel
MIT Sloan School of Management
Cambridge, MA 02142, USA
{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. We propose a solution based on extensions to the Context Interchange framework, which has 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 with 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 counterterrorism efforts. It is often even required by law for organizations to store historical data and make sure it is accurate and easy to retrieve. 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 historical database, see Figure 1. Two anomalies 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!

1 See Robert Shier on "Regulated storage" in ComputerWorld, 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.

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

These anomalies 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.

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. Temporal database research has primarily focused on management of temporal data in a homogeneous context environment. 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 techniques to effectively resolve temporal semantic conflicts between data sources and receivers. Specifically, we extend the Context Interchange...
(COIN) framework [5, 9] with temporal contextual knowledge representation and reasoning capability.

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 [12]. 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.

<table>
<thead>
<tr>
<th>Year</th>
<th>Num_Employee</th>
<th>Profit</th>
<th>Tax</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>5100</td>
<td>4.2</td>
<td>1.1</td>
</tr>
<tr>
<td>2000</td>
<td>12000</td>
<td>13000</td>
<td>2500</td>
</tr>
<tr>
<td>2001</td>
<td>25.3</td>
<td>20000</td>
<td>4800</td>
</tr>
<tr>
<td>2002</td>
<td>30.6</td>
<td>35.3</td>
<td>7.97</td>
</tr>
</tbody>
</table>

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>
<thead>
<tr>
<th>Currency</th>
<th>Source</th>
<th>Receiver</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Francs(FRF), year ≤ 2000</td>
<td>USD, always</td>
</tr>
<tr>
<td></td>
<td>Euros, year ≥ 2001</td>
<td></td>
</tr>
<tr>
<td>Scale factor for profit and tax</td>
<td>1M, year &lt;= 1999</td>
<td>1K, always</td>
</tr>
<tr>
<td></td>
<td>1K, 2000 ≤ year &lt; 2001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1M, year &gt;= 2002</td>
<td></td>
</tr>
<tr>
<td>Scale factor for Num_Employee</td>
<td>1K, year &lt;= 2001</td>
<td>1K, always</td>
</tr>
<tr>
<td></td>
<td>1K, year &gt; 2002</td>
<td></td>
</tr>
<tr>
<td>Profit</td>
<td>Exclude tax, year &lt;= 2000</td>
<td>Include tax, year &lt;= 2001</td>
</tr>
<tr>
<td></td>
<td>Include tax, year &gt; 2001</td>
<td>Include tax, always</td>
</tr>
</tbody>
</table>

Table 2 gives examples of the context knowledge in a simple integration scenario involving the source in Table 1 and a 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 scenarios that involve dozens of sources and receivers, each with time varying heterogeneous contexts. We need effective technologies to manage this complexity.

2.2. Temporal semantic heterogeneities

We see at least three categories of issues in the integration of temporal data.

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. 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" used to refer to Chrysler but changed to refer to Citigroup on December 4, 1998 after Chrysler merged with Daimler-Benz. The derivation method of concepts, such as unemployment rate, often change over time.

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 [21]. 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 system service to record con-
text and use it 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 each provides useful concepts and techniques, none address all the 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.

Temporal databases. 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 may have a duration and thus is not an instant in time ontology [10].

Over 40 temporal data models have been proposed [20]. 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 [4]. 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 data warehouses. This ad-hoc approach lacks flexibility and scalability.

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 [16] 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 [11] can be applied. We will use a solver implemented using constraint handling rules (CHR) [8] as demonstrated in [6].

Temporal granularity research, both logic [17] 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.

Data integration. Approaches to achieving data integration largely fall into tight-coupling and loose-coupling categories depending on whether a global schema is used [5, 9]. 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 [14] in rewriting user queries for the former approach and the user's burden of writing complex queries for the latter limit the flexibility and scalability of these approaches.

COIN [5, 9] is a middle ground approach that, avoids these shortcomings by encapsulating data semantics into a context theory and maintaining accessibility of source schema by users. In COIN, a user issues queries as if all sources are in the user's context and a mediator is used to automatically rewrite the queries to resolve semantic differences.

Unfortunately, existing approaches, including COIN, use a static data model and ignore temporal context. Consequently, temporal concepts are missing in the ontologies used in these systems. This research will focus on the representation and reasoning of temporal context. Our framework incorporates context into the 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 COIN demonstrates its new capability in solving ontological heterogeneity problems [5]. With various temporal representation and reasoning techniques, we can further extend COIN to handle temporal semantic heterogeneities.

4.1. The COIN framework

The COIN framework [5] 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 (ACLPL) [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 types that correspond to the domain and the range of the attribute.

The ontology is augmented with temporal concepts as defined in the time ontology [10]. The most general one is 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 subtype. 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. A regular attribute represents a relation 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, defined by context, functionally determines the interpretation and the value of the regular attribute [20]. For example, currency in Table 1 is a contextual attribute 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. We achieve 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 in user's context, the mediator detects and reconciles semantic differences by examining involved contextual attributes and applying appropriate conversion functions if the values differ 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 MQ can be generated as in the case of existing COIN; the interval appears in the 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 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 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 and mediation

Representation. Figure 2 shows the ontology for the example. Here, we use the most general concept temporal entity. Using elevation axioms, we create the mappings between attributes in data source and the types in the ontology.

![Figure 2. Example ontology and elevation](image-url)
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. We implement the ≤ relation between points as a 'tile' constraint. The contains relation between an interval and a point is translated into tile constraints; the overlaps relation between intervals are also translated into tile constraints.

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

```prolog
modifier(moneyValue,0, currency, c_src, M) :-
    containObj([bottom, 2000], O),
    case(basic, N, c_src, "FRF"),
    modifier(moneyValue, 0, currency, c_src, M) :-
    containObj([2002, top], O),
    case(basic, N, c_src, "EUR").
```

The head of the statement reads: O of type moneyValue has a contextual attribute (i.e., modifier) currency, whose value in source context c_src is M. Its body has two predicates. containObj([1, O]) will use the tense of O to obtain its temporal attribute T (which corresponds to the attribute in the EDB) of type temporalEntity and add constraint contains(T, T). The helper predicate case specifies the primitive value of T in c_src context. Thus, the history of each contextual attribute is now a set of pairs \( \langle \nu_i, 1 \rangle \), where \( \bigcup_i \nu_i = \text{[bottom, top]} \).

For contexts that do 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.

**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:

\[
\text{[<moneyValue, currency, "FRF">,}
\text{<moneyValue, scale, "1000000">,}
\text{<moneyValue, type, "taxExcluded">,}
\text{<sem_num, scale, "1"}>].
\]

![Figure 3. Visualization of temporal context](image)

Recall that we translate all temporal relations into tile constraint over points. Each contextual attribute will generate two tile constraints for the temporal variable. The above problem is thus turned into 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 abduced predicates. Constraints over other time points can be pair-wise simplified. We also implement overlaps to simplify tile 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 meet-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 allowing us to reuse the non-temporal implementation of the COIN mediator.

### 5.2. Preliminary results

These temporal extensions to COIN framework enable semantic interoperability for the integration example. The prototype can generate SQLs 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 using 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. 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):

```
\text{\$1, \text{\$1}-1999, currency, scale, type, scale}
answer(\text{\$24}, \text{\$23}, \text{\$22}, \text{\$21});
\text{\$24} \equiv \text{\$23} \equiv \text{\$22} \equiv \text{\$21} \equiv \text{\$20};
\text{\$20} \equiv \text{\$19} \equiv \text{\$18} \equiv \text{\$17} \equiv \text{\$16} \equiv \text{\$15} \equiv \text{\$14} \equiv \text{\$13} \equiv \text{\$12} \equiv \text{\$11} \equiv \text{\$10} \equiv \text{\$9} \equiv \text{\$8} \equiv \text{\$7} \equiv \text{\$6} \equiv \text{\$5} \equiv \text{\$4} \equiv \text{\$3} \equiv \text{\$2} \equiv \text{\$1} \equiv \text{\$0};
\text{\$21} \equiv \text{\$20} \equiv \text{\$19} \equiv \text{\$18} \equiv \text{\$17} \equiv \text{\$16} \equiv \text{\$15} \equiv \text{\$14} \equiv \text{\$13} \equiv \text{\$12} \equiv \text{\$11} \equiv \text{\$10} \equiv \text{\$9} \equiv \text{\$8} \equiv \text{\$7} \equiv \text{\$6} \equiv \text{\$5} \equiv \text{\$4} \equiv \text{\$3} \equiv \text{\$2} \equiv \text{\$1} \equiv \text{\$0};
\text{\$22} \equiv \text{\$23} \equiv \text{\$24};
\text{\$23} \equiv \text{\$24};
\text{\$24};
```

113
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. These subqueries resolve all the semantic conflicts in Table 2 or in Figure 3.

6. Discussion and future plan

We identified three types of semantic heterogeneity in the integration of temporal data. There is an ever increasing need to efficiently handle temporal heterogeneities as more historical data is used for auditing, 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 [7]. We need to investigate how this logic can be integrated with the ACLP based COIN mediator.

An important part of future research will focus on the heterogeneities of temporal entities. We plan to add various contextual attributes to the temporal entity type in the ontology and use 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 of time and date can have serious consequences, as has shown in an 1805 event [15] 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

PROCEEDINGS

11th International Symposium on TEMPORAL REPRESENTATION AND REASONING
TIME 2004

1-3 July 2004, Tatihou Island - Normandie - France

Edited by Carlo Combi and Gérard Ligozat

Sponsored by