CISL: Composing Answers from Disparate Information Systems

Stuart E. Madnick, Y. Richard Wang, Meryl Alford, Alex Champlin,
Francis Gan, Howard Gerber, David Godes, Robert Goldberg,
Amar Gupta, Dave Horton, Rebecca Kao, Sam Levine, Michael Madnick,
Bryan McCay, Leslie McCafferty, Allen Moulton, Mia Paget,
Bertrand Rigaldies, Michael Siegel, Terence Sim,
T.K. Wong, and Yeuk Yuan.

September 1989 WP# 3091-89-MS

CISL:
Composing Answers from Disparate Information systems

Stuart E. Madnick *
Y. Richard Wang*
Meryl Alford
Alex Champlin
Francis Gan
Howard Gerber
David Godes
Robert Goldberg
Amar Gupta
Dave Horton
Rebecca Kao
Sam Levine
Michael Madnick
Bryan McCay
Leslie McCafferty
Allen Moulton
Mia Paget
Bertrand Rigaldi
Michael Siegel
Terence Sim
T.K. Wong
Yeuk Yuan

* Principal investigators of the CISL research project.

September 1989

Composite Information Systems Laboratory
Room E53-320, Sloan School of Management
Massachusetts Institute of Technology
Cambridge, Massachusetts, 02139
ATTN: Prof. Y. Richard Wang
(617) 253-0442 rwang@sloan.mit.edu

Work reported herein has been supported, in part, by Citibank, IBM, Reuters, MIT's International Financial Service Research Center (IFSRC), MIT's Laboratory of Computer Science (LCS), and MIT's Leaders for Manufacturing (LFM) program.

Extended Abstract

MOTIVATION

Significant advances in information technologies have provided opportunities for dramatically increased connectivity among information systems. Meanwhile, globalization, whereby the scope and presence of organizations expand beyond their traditional geographic boundaries, has exerted propelling effect on many organizations to capitalize on these increased opportunities for connectivity. As a result, many important applications in the 1990's will require access to and integration of multiple disparate databases both within and across organizational boundaries. It has also become increasingly clear that a top-down data integration approach is not appropriate for this type of applications.

CISL RESEARCH

This paper presents the problems we experienced and solutions developed in the CISL research project[1] for composing answers from disparate information systems. A fundamental CISL assumption is that organizations must deal with pre-existing information systems which have been developed and administered independently, and are likely to remain so. With this assumption, CISL follows two principles: (1) *system non-intrusiveness*, and (2) *data non-intrusiveness*.

By *system non-intrusiveness* we mean that no CISL system components should need to be added to a pre-existing information system. A system-intrusive example, arguably, is MULTIBASE [Smith et al, 1981; Goldhirsch et al, 1984] in which an LDI is installed at each pre-existing information system site. In our experiences, we have found that many of these systems are controlled by autonomous organizations or even separate corporations (e.g., Dow Jones financial services) that at best would be reluctant or slow to change their systems and at worst would not be willing to do so. By *data non-intrusiveness* we mean that no data should need to be altered in a pre-existing information system in order to allow for data integration. A data-intrusive example is where data in pre-existing information systems must be changed to a standardized form. Although we are in favor of data standardization, in our studies we have found it difficult to attain in a timely manner across organizational boundaries. *To a pre-existing information system, in short, CISL will behave as a regular user.*

A primary CISL focus is logical connectivity[2] which is concerned with the ability to know where and how data are stored, to decompose a user query into sub-queries that can be executed by local systems,

---

1 The CISL project is being conducted at the Composite Information Systems Laboratory, Sloan School of Management, MIT. Recognizing the criticality of organizational autonomy and strategic deployment of information technologies, the CISL research effort has addressed four related aspects of connectivity: strategic, organizational, physical, and logical [Wang and Madnick, 1988]. Furthermore, a CISL prototype has been implemented. Currently, the system allows for simultaneous access to relational, menu-driven and command-driven databases in multiple machines.

2 Also referred to as (intelligent) *interoperability* [Manola, 1988; 1989].

to accumulate the results from all the sub-queries, to reconcile differences among the results accumulated, and to formulate composite answers for the user. In particular, we are interested in (1) *semantic reconciliation* which deals with the integration of data semantics among disparate information systems, and (2) *dynamic query composition* where new information is inferred and composed based on the underlying data in a heterogeneous distributed environment. This requires CISL to *explicitly represent more of the semantics in each local database and the semantic heterogeneities across disparate information systems.* Specifically, we are seeking a semantic representation which will allow us to capture what a view administrator[3] knows about the local databases explicitly. The more semantic heterogeneities we can represent in CISL, the more we can automate our *semantic reconciliation* and *dynamic query composition* activities so that CISL will behave more intelligently in dealing with unexpected ad hoc queries.

Related Research

The CISL project has benefited from other research efforts on heterogeneous database systems design and implementation, notably MULTIBASE in the U.S., PRECI* in England [Deen, Amin, and Taylor, 1987a, 1987b], and MRDSM in France [Litwin and Abdellatif, 1986]. To assess the state-of-the-art commercial heterogeneous distributed DBMS capabilities, we are also surveying commercial systems offering solutions to the distributed heterogeneous database problem, for example Cincom's SUPRA, Metaphor's DIS, Oracle's SQL*loader, and TRW's Data Integration Engine (DIE).

In terms of *semantic reconciliation* and *dynamic query composition*, most of the commercial systems provide ad hoc[4] solutions. Although ad hoc solutions may be sufficient for a particular application in the short run, they hinder the representation of the semantics in disparate information systems A more systematic approach to the semantic heterogeneity problem is through schema integration. To our knowledge, schema integration and the subsequent semantic reconciliation tasks have been accomplished either by a view administrator[5] or a user given his understanding of the export schemas.[6]

Hull and King [1987] summarized two philosophical approaches to semantic modeling: entity (e.g., ER and RM/T) vs. attribute-based (e.g., FDM). In developing a semantic representation scheme for CISL, we have realized that it is critical to have the strengths of both approaches. The issue here is how to provide a theory-based representation scheme. We are investigating the possibility of a combination of the extended ER model [Teorey, Yang, and Fry; 1986; Batini, Lenzerini, and Navathe, 1986], the RM/T model [Codd, 1979], and ADT [Weller and York, 1984; Stonebraker, 1986] with a formal representation

---

3 in the Multibase sense.

4 For example, UNIX shell scripts have been used in some systems to resolve data incompatibilities and semantic mismatches. This approach opaques the semantic heterogeneity and prohibits semantic information from being shared.

5 As in the Multibase project where a view administrator is authorized by multiple DBAs to interpret their database schemata.

6 As in the MRDSM and PRECI* projects.

which will include the fundamental entities, attributes, sub/super types, domains, functions, and view definition. The formal representation, in turn, will allow us to further investigate, among others: (1) how *semantic reconciliation* and *dynamic query composition* may be accomplished given the formal representation at the local and composite level; and (2) the conditions under which a new local database can be incorporated without revising the composite model; and if revisions are required, the scope and algorithms.

We now turn our attention to synopsize the current CISL database environment, followed by an example to illustrate some interesting CISL problems and approaches.

CISL DATABASE ENVIRONMENT

The current CISL has access to three MIT databases (the alumni database, the Sloan recruiting database, and the Sloan student database) and three commercial databases (Finsbury's **Dataline** and I.P. Sharp's **Disclosure** and **Currency**[7]). These databases provide breadth in data and provide examples of differences in style, accentuated somewhat by the different origins of each service -- Finsbury is based in London, England and I.P. Sharp is based in Toronto, Canada.

The three MIT databases are all dialects of SQL running on different computers operated by different MIT organizations: alumni database (Informix-SQL on an AT&T 3B2 computer), recruiting database (Oracle-SQL on an IBM PC/RT), and the student database (SQL/DS on an IBM 4381).

Dataline provides financial statements from the previous five years and forecasting facilities for over 3000 corporations. The data cover primarily U.K. corporations as well as a number of other major international firms. Queries to **Dataline** are menu-driven and results are returned in set formats.The database system appears hierarchic to the end-user -- company records can only be accessed through a company code which the user must look up in a published booklet or through an on-line company-name/company-code table. **Dataline** is accessible from the United States via Telenet.

Data in **Disclosure** are obtained from financial reports of over 12,000 companies which report to the Securities Exchange Commission of the U.S.[8] The companies are primarily incorporated in the US, but a number of non-US companies are also included. Users can formulate queries based on company names, codes, ticker symbols or CUSIP numbers, or queries can be made based on values for any static or time series data. The user can access **Disclosure** through menus or using a proprietary query language available on the I.P. Sharp communications network. Data can be returned in predetermined formats, user-specified formats, or as a stream of data for downloading onto a computer.

Currency contains historical daily records of exchange rates of currencies traded on world markets. Exchange rates are reported for several major exchange market locations (e.g., New York, London,

---

7 Finsbury Data Services and I.P. Sharp Associates are both owned by Reuters Holdings PLC.

8 Disclosure is produced by Disclosure Incorporated, Bethesda, Maryland, USA and is distributed through a number of information services.

Singapore, and Japan) in each case based on the local currency of that market. Data contained in **Currency** are obtained from major banks or financial institutions in the countries where the markets are located. Like **Disclosure** and other databases available through I.P. Sharp, **Currency** can be accessed via menus or query languages, and output can be obtained in tabulated on stream formats [Paget, 1989].

CISL EXAMPLE AND APPROACH

We now present an example to help illustrate both the problems to be solved and approaches used in CISL. The example is a simplified version of actual heterogeneous environments that we have analyzed [Paget, 1989; Godes, 1989].

Consider the following two customer databases, one for I.P. Sharp and the other for Finsbury Data Services, as depicted in Table I and II respectively. Reuters has recently purchased both corporations as part of its movement into providing historical financial data services. Since both corporations were, and to a large extent still are, operated separately, so are their customer databases. Suppose, as part of a major new marketing campaign, Reuters' managing director wants to *know the top hundred customers in terms of total purchases of historical financial information services, as provided by I.P. Sharp and Finsbury, over the past five years, expressed in dollars*.

Table I: A Simplified I.P. Sharp Customer Database

Company	CEO	Purchase	Year	City
IBM Corp.	John Ackers	$60,000.	1988	
MIT	Paul Gray			Cambridge
Continental Airlines				
Toyota		¥120,000		
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

Table II: A Simplified Finsbury Customer Database

Customer	CEO	Purchase (000's)	Country	FY	HQ
IBM, Inc.		50	USA	311288	
Mass. Inst. Tech.	Paul Gray		USA		Cambridge
Texas Air Corp			USA		
Honda		210	Japan		
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

Total purchases for a customer is calculated by summing purchases from each of the two services (I.P. Sharp and Finsbury). In order to *reconcile the semantic heterogeneities*, as discussed below, two major tasks need to be addressed by the *dynamic query composer*: (a) instance matching and (b) value interpretation and coercion [Wang and Madnick, 1989a, 1989b].

Instance Matching

As illustrated in Table I and II, each service may use a different spelling of the name for a customer. Thus, it will be necessary to match a customer in one database to the same customer in the other by means other than an exact string comparison. For example, "IBM Corp." in I.P. Sharp should be matched to "IBM, Inc." in Finsbury; also "MIT" should match "Mass. Inst. Tech."; and "Continental Airlines" should match "Texas Air Corp.". Each service may also use a customer number (or other unique identifier) for a customer, but it is unlikely that the same identifier would have been used by both services for a given customer.

To match customers (or any entity) between two databases, we use a combination of three techniques: key semantic matching, attribute semantic matching, and organizational affinity.

- *Key Semantic Matching*: to match the two IBM's, we can use rules such as "Corp." and "Inc" suffixes are equivalent. These rules can be context dependent. As a simple example, we would not want "Inc Magazine" to be matched with "Corp Magazine".
- *Attribute Semantic Matching:*. The two MIT's are harder cases since it is unlikely that we would have a rule that "M" and "Mass." were equivalent. Instead we identify the match based upon the attributes (both have CEO "Paul Grey" and HQ city "Cambridge").

After semantic instance matching (both key and attribute) has been performed, the results can be saved in the Inter-Database Instance Identification Table (IDIIT). Thus, subsequent join operations can be accomplished rapidly through automatic use of the IDIIT.

- *Organizational Affinity*: In many cases there exists an affinity between two distinct organizations. For example, the "IBM Federal Systems Division" is distinct from "IBM Corp" (i.e., it is a division of IBM Corp). Also, although "Continental Airlines" is a separate corporation from "Texas Air Corp", Texas Air Corp owns Continental. Thus, depending on the purpose of the query, it may be desirable to treat two distinct entities as being the same (e.g., from a marketing perspective Continental and Texas Air may be merged, from a legal liability perspective they should be kept separate). The organizational affinity facility requires two steps: First, external knowledge must be supplied to describe the nature of the affinity between the organizations -- this may be supplied manually or through processing of other organizational databases. This information is held within the Inter-Database Instance Affinity Table (IDIAT). Second, a context can be provided for any query or session (e.g., "if entity A owns more than 51% of entity B, then treat them as matching entities") which will be used to define generalization in the instance matching process.

Value Interpretation and Coercion

After instance matching is performed, we must compute the combined sales to each customer. This presents numerous challenges regarding value interpretation and value coercion.

- *Value Interpretation:* in the I.P. Sharp database, the total purchases is expressed as a character string that combines the currency indicator and numeric value (e.g., ¥120,000). This information must be interpreted. Finsbury does not explicitly identify the currency being used; it must be inferred from the country information (USA = $, Japan = ¥). Furthermore, Finsbury stores purchase amount in thousands of whatever currency is applicable to the country of the customer. For example, Honda's purchases are ¥210,000, but IBM's are $50,000.
- *Value Coercion:* In order to compute the total purchase in dollars (or to compare totals across customers), a currency exchange rate needs to be applied. Since the yen/dollar exchange rate varies considerably from year to year (if not moment to moment), it is also necessary to "know" where to find the currency exchange table for the time period covered by the data used. I.P. Sharp stores the "Year" as "1988", while Finsbury stores "FY" as "311288", requiring yet another conversion to be performed.

Source of Data

Finally, users may want to know the source of the data presented (e.g., "source: Reuters' Newstext, Sept. 22, 1989"), so that they can apply their own judgement on the quality of the information. We are developing a data-tagging mechanism (called *polygen relational algebra)* to tag sources of information (for each cell, tuple, attribute, and relation) from the local level, to the composite level, and to the user's composite answers. Upon request, the sources can be displayed, and further composite answers can be formulated based on preferences indicated by the user.

CURRENT CISL STATUS AND DIRECTIONS

CISL Status

CISL version 3.0 was completed in August 1989. It runs on an AT&T 3B2/500 computer under UNIX System V. It is mostly implemented in our object-oriented rule-based extension of Common Lisp, called the Knowledge-Oriented Representation Language (KOREL). Certain components are implemented in C and UNIX Shell.

The current system provides direct access to, and integration of, information from all six databases described earlier. Many of the instance matching, value interpretation and coercion features are in rudimentary form at the present time.

CISL Directions

This paper has only highlighted a portion of the CISL goals. The directions can be summarized into three categories: (1) more robust and comprehensive designs and implementations of the features described here, (2) continuation of the design and implementation of unmentioned features (e.g., synchronization of changes in data semantics between systems), and (3) continuation of our surveys and analyses of the strategic and organizational factors that impact the heterogeneous database environment.

CONCLUDING REMARKS

We have presented the CISL research project's efforts in logical connectivity. Recent business changes are both enabled by and are the driving forces towards increased connectivity. Some of the increasingly evident changes include innovative strategic information systems requiring a high level of cross-functional integration such as airline reservation systems locking in independent travel agents, direct order entry and order status inquiry systems between buyers and suppliers, and corporations linked in information networks for global security trading. We believe that corporations well positioned for increased connectivity will have a competitive advantage in the decade ahead. The key to increased connectivity is the capability to develop heterogeneous database solutions in general, and logical connectivity in particular.

REFERENCES

Batini, C., Lenzerini, M., and Navathe, S.B. [1986] "A comparative analysis of methodologies for database schema integration." *ACM Computing Surveys*, Vol. 18, No. 4, pp. 323-364.

Codd, E. F. [1979]. "Extending the Relational Model to Capture More Meaning." in *the ACM Transactions of Data Base Systems.*

Dayal, U., Hwang, H., Manola, F., Rosenthal, A.S., and Smith, J.M. [1984] "Knowledge-oriented database management: Final technical report, Phase I ." Computer Corporation of America, Cambridge, MA .

Deen, S.M., Amin, R.R., and Taylor, M.C. [1987a] "Data integration in distributed databases." *IEEE Transactions on Software Engineering*, Vol. SE-13, No. 7, pp. 860-864.

Deen, S.M., Amin, R.R., and Taylor, M.C. [1987b] "Implementation of a prototype for PRECI*" *Computer Journal*, Vol. 30, No. 2, pp. 157-162.

DeMichiel, L.G. [1989] "Performing operations over mismatched domains." *Proceedings of the Fifth International Conference on Data Engineering*, Los Angeles, CA, February 1989.

Godes, D.B. [1989] "Use of Heterogeneous Data Sources: Three Case Studies." Sloan School of Management, MIT, Cambridge, MA. CISL Project, WP # CIS-89-02.

Goldhirsch, D., Landers, T., Rosenberg, R., and Yedwab, L. [1984] "MULTIBASE System Administrator's Guide," Computer Corporation of America, November.

Heimbigner, D. and McLeod, D. [1985] "A federated architecture for information management." *ACM Transactions on Office Information Systems*, Vol. 3, No. 3, pp. 253-278.

Hull, R. and King, R. [1987] "Semantic database modeling: Survey, applications, and research issues." *ACM Computing Surveys*, September 1987, Vol. 19, No. 3,

Litwin, W. and Abdellatif, A. [1986] "Multidatabase interoperability." *IEEE Computer*, December, p. 10-18.

Manola, F. [1988] "Distributed object management technology." Technical Memorandum, GTE Laboratories, TM-0014-06-88-165.

Manola, F. [1989] "Applications of object-oriented database technology in knowledge-based integrated information systems." Paper prepared for the CRAI School on Recent Techniques for Integrating Heterogeneous Databases, Venezia University, April 10-14.

McLeod, D.J. [1976] "High level domain definition in a relational data base system." IBM Research Laboratory, San Jose, CA.

Paget, M.L. [1989] "A knowledge-based approach toward integrating international on-line databases." Sloan School of Management, MIT, Cambridge, MA. CISL Project, WP # CIS-89-01.

Peckham, J. and Maryanski, F. [1988] "Semantic data models." *ACM Computing Surveys*, Vol. 20, No. 3, pp. 153-189.

Qian, X. and Wiederhold, G. [1986] "Knowledge-based integrity constraint validation." *Proceedings of the Twelfth International Conference on Very Large Data Bases*, pp. 3-12.

Rusinkiewicz, M., Emasri, R., Czejdo, B., Georgakopoulos, D., Karabatis, G., Jamoussi, A., Loa, K., Li, Y., Gilbert, J., and Musgrove, R. [1988] "Query processing in omnibase - a loosely coupled multi-database system." University of Houston, Technical Report #UH-CS-88-05.

Smith, J. M., Bernstein, P.A., Dayal, U., Goodman, N., Landers, T., Lin, K.W.T., and Wong, E. [1981] "Multibase - Integrating heterogeneous distributed database systems." 1981 National Computer Conference, pp. 487-499.

Stonebraker, M. [1986] "Inclusion of new types in relational data base systems." *IEEE*, pp. 480-487.

Teorey, T.J., Yang, D., and Fry, J.P. [1986] "A logical design methodology for relational databases using the extended entity-relationship model." *Computing Surveys*, Vol. 18, No. 2, pp. 197-222.

Wang, R. and Madnick, S. [1988] *Connectivity among information systems.* Composite Information Systems (CIS) Project, Vol. 1, 141 pages.

Wang, R. and Madnick, S. [1989a] "Facilitating *Connectivity in Composite information systems,"* To appear in *the ACM, Database.*

Wang, R. and Madnick, S. [1989b] "The Inter-Database Instance Identification Problem in Integrating Autonomous Systems," *Proceedings of the Fifth International Conference on Data Engineering*, February 6-10, 1989.

Weller D.L., and York, B.W. [1984] "A relational representation of an abstract type system." *IEEE Transactions on Software Engineering*, May, Vol. SE-10, No. 3, pp. 303-309.

Wong, T.K. [1989] "Data connectivity for the Composite Information System/Tool Kit." Sloan School of Management, MIT, Cambridge, MA. CISL Project, WP # CIS-89-03.