The Model-Assisted Global Query System For Multiple Databases in Distributed Enterprises

by

Waiman Cheung [1] and Cheng Hsu [2]

October 1993

Revised July 1994

Submitted to ACM Transactions on Information Systems

[1] **Lecturer, Faculty of Business Administration, The Chinese University of Hong Kong, Shatin, Hong Kong. email: wcheung@cuhk.hk**

[2] **Associate Professor, Decision Sciences and Engineering Systems, Rensselaer Polytechnic Institute, Troy, N.Y. 12180-3590. email: hsuc@rpi.edu**

ABSTRACT

Today's enterprises typically employ multiple information systems, which are independently developed, locally administered, and different in logical or physical designs. Therefore, a fundamental challenge in enterprise information management is the sharing of information for enterprise users across organizational boundaries; which requires a global query system capable of providing on-line intelligent assistance to users. Conventional technologies, such as schema-based query languages and hard-coded schema integration are not sufficient to solve this problem.

This research develops a new approach, "model-assisted global query system," that utilizes an on-line repository of enterprise metadata - the metadatabase - to facilitate global query formulation and processing with certain desirable properties such as adaptiveness and open systems architecture. A definitional model characterizing the various classes and roles of the minimally required metadata as knowledge for the system is presented. The significance of possessing this knowledge (via a metadatabase) towards improving the global query capabilities available previously is analyzed. On this basis, a direct method using model traversal and a query language using global model constructs are developed along with other new methods required for this approach. It is then tested through a prototype system in a Computer-Integrated Manufacturing setting.

Categories and Subject Descriptors: H.2.1 [Database Management]: Languages-Query Language; H2.4 [Database Management]: Systems-Distributed Systems, Query Processing; H2.7 [Database Management]: Database Administration-Data Dictionary/Directory; H3.3 [Information Storage and Retrieval]: Information Search and Retrieval-Query Formulation; H5.2 [Information Interfaces and Presentation]: User Interface

General Terms: Information Systems Integration, Interoperability, Metadata

Additional Key Words and Phrases: Heterogeneous Distributed Information Systems, Global Query System, Information Sharing

Acknowledgment

This Research is supported in part by National Science Foundation Grants DDM 9015277 and DDM 9215620, and Rensselaer's CIM and AIME Programs, both of which are sponsored by ALCOA, DIGITAL, GE, CM, and IBM and conducted under the Center for Manufacturing Productivity and Technology Transfer.

1. GLOBAL QUERY SYSTEMS

The notion of information-based enterprises have become a reality in '90. For various organizational and technological reasons, information systems in these enterprises are typically characterized by heterogeneous, distributed environments. For example, a computerized manufacturing enterprise may have a number of shop floor sub-systems that are implemented with different file management systems on various platforms, while its business and engineering design sub-systems are operating in relational and object-oriented environments distributed over wide-area networks. In all likelihood, this multiplicity will not disappear, nor replaced by an all-encompassing standard any time soon. Therefore, a major objective of information integration for these enterprises is to provide a logical structure to integrate these islands of information resources for enterprise-wide information sharing and management without relying on a fixed controlling hierarchy. A key requirement here is what might be called thon-line intelligence and assistance" capabilities of the integrated systems for supporting enterprise users' (varying) needs in retrieving information at a global level from the multiple local systems, which may frequently change their operating rules as well as contents. Numerous research and commercial systems have evolved over the past decade towards providing these capabilities for single-site or multiple-site databases. However, a rigorous formulation of this requirement for expressly the global query needs of multiple systems has not been provided previously in the literature; nor has such a technology.

The Need for Enterprise Metadata Support

Metadata has been increasingly recognized as a key element in global query systems [4, 9, 12, 14, 16, 28, 29, 49, 54]. The question raised here is, how much and what metadata a global query system should process in order to effect the on-line intelligence and assistance capabilities, and how to develop an architecture which manages and utilizes the metadata to suffice the end?

To illustrate the significance of these envisioned capabilities, we consider below some basic tasks required of a global query system in a heterogeneous, distributed environment (see, e.g., [12, 34, 36, 42] for a survey on these tasks). A typical global query operation involves two steps; namely, global query formulation and global query processing. In the first step, the user's requests are articulated and represented in a way that the global query system understands. The query formulation is done primarily through the user interface of the system. In the second step, accomplished by the system internally, queries are sent to appropriate local systems to retrieve pertinent information and reassembled for the users.

Towards query formulation, on-line intelligence enables the user interface to provide assistance in the articulation as well as allow for high-level, intuitive representation of queries. Specifically, for the articulation of queries, the system would utilize its knowledge on the enterprise information resources (which are referred to in this paper as metadata; including information models, implementation models and business and operating rules) to alleviate semantic ambiguity, facilitate logical analysis, and enhance adaptive construction during the formulation of queries. Similarly, the representation itself could accommodate heterogeneity in

local models across the enterprise through, e.g., the knowledge on enterprise SUBJECTs and the equivalence of data items among different local systems without having to impose a single, fixed "integrated schema" on all databases.

In addition to supporting high-level and non-syntax-based queries for enterprise users, the system would also handle all context-based interpretations and dynamic mappings between the globally formulated query and the locally implemented file structures or database schemata. Assisted with these capabilities, the second step - global query processing - would be accomplished in the following fashion:

(1) query optimization: the global query is first optimally decomposed into local queries taking into account both semantics and performance;

(2) query translation: the local queries are then encoded in their respective data manipulation languages;

(3) query execution: the encoded local queries get dispatched to and processed at their respective systems; and finally

(4) result integration: results from local systems are assembled to answer the global query.

Each of these processing steps makes use of metadata as well. For instance, local database schemata, directory and network information, and the contextual knowledge of data are required for query optimization; local DBMS information for query translation; operating rules on information flows for query execution; and knowledge on equivalent objects, incompatibility and data conversion for result integration.

Without the assistance from sufficient on-line knowledge in the form of metadata, all of the remaining information required above for both

query formulation and processing would have to be provided by the users or application programs, such as the case in pervious systems regardless of the interfacing designs used [1, 2, 3, 6, 13, 19, 38, 47, 48, 51, 54, 55]. The problem of lacking on-line metadata support is especially acute for multiple systems, where researchers have increasingly emphasized the use of enterprise metadata (see, e.g., [9, 28] for a survey on this topic). A basic reason is the significant differences in the CONTEXTs in which data are utilized, on top of the complexities in the varying data semantics (models), data manipulation languages, and data structures among local systems (a discussion of the contexts is provided in [27, 30]). Either the users or the global system itself must abridge these differences for each query before the information can be shared. Since enterprise users generally do not possess the expertise in database technology, nor the technical knowledge about the local systems, they cannot truly benefit from a global query system which does not possess sufficient metadata to provide on-line intelligence and assistance. The notion of enterprise metadata is concerned with all metadata pertaining to the above SUBJECTs and CONTEXTs.

User Interface for Query Formulation

User interface techniques are important to global query systems; however, do they supplant the need for enterprise metadata support? Techniques such as windows, icons, menus, graphics, visualization [19, 20, 21, 39, 45] and other forms of non-textual formalisms (see, e.g., [32, 37] for a survey), free the typical non-programmer user from having to learn sophisticated programming languages. Therefore, graphical user interface (GUI) technologies have been employed together with cognition-theoretic interface design principles and guidelines [18, 19] to facilitate database

query tasks. The results have improved significantly the commercially available database query systems, especially those emerging in the '90s (e.g., GUI add-ons to SQL for a few relational systems [5]). Notwithstanding, these latest products still do not support users with on-line metadata on enterprise models (especially contextual knowledge of data) required for global query system in heterogeneous environments. Users are still charged with the responsibility of furnishing many of the technical information mentioned above. Moreover, since these systems do not offer an on-line global model separate from the schemata, which are fixed, their user interfaces tend to be hard to change or to customize, as the underlying systems or the users change.

The same observations are largely applicable to the so-called natural language interfaces. In principle, any systems that use natural languages should not require the users to learn the artificialities of correct command formats or modes of interaction [8, 10, 23, 40, 44]. Unfortunately, few systems (even research systems) have successfully achieved this goal, due mainly to insufficient results in linguistics and artificial intelligence to support this class of user interface. However, even in the ideal case, a natural language interface would not be able to assist on the query formulation itself for the same reasons as GUI faces: It does not provide users with such knowledge as the local and global data resources, dictionary and directory knowledge, and business rules about the enterprise; all of which are needed before a user interface could ever support non-technical user in a "natural" mode of query formulation.

Addressing the need of providing metadata to users, database browsers have been developed to help end users to look through the contents of a particular database. Most of these database interfaces are

limited to browsing the data instances and support only single-site and single-tuple queries; moreover, few look beyond the simple database schema per se, which does not include other types of enterprise metadata [22, 38, 50, 52,56]. They, ironically, provide interesting evidence supporting the thesis that enterprise metadata can lead to a new breakthrough for the problem.

Integrated Schema for Global Query

Enterprise metadata as discussed above are complex in its own right. Thus, it is natural to expect an architecture devoted to them for the above stated tasks. Is the conventional integrated schema technology sufficient?

A key issue concerning expressly multiple systems is how to reconcile and consolidate the various views and representation methods across the enterprise and yet still retain local differences for autonomy and flexibility. Most of previous efforts employ a solution strategy emphasizing the development of global architectures based on the schema integration (e.g., [7]) approach. Although an integrated schema is a facility of metadata, its the hard-coded nature tends to contradict or even nullify some of the basic promises of true local autonomy such as openness and adaptiveness [36]. Moreover, research efforts (e.g., [1, 13, 34, 49, 54]) have also revealed that additional enterprise metadata beyond the integrate schema are needed to facilitate the representation of global views and the management of query transactions amount local databases. Therefore, these efforts have by contradiction shown an even bigger role for enterprise metadata; that is, minimizing the reliance on some fixed, hard-coded global schemata or controller to effect information integration.

Thus, the conventional approach of integrated schema does not seem to provide a solution to the enterprise metadata management problem. To bring the point into better light for our discussion, we refer to the concept of an independent system of enterprise metadata supporting the above purposes as a *metadatabase*.

The Objective and Organization of the Paper

A conclusion from the above discussion is evident: the man-machine interface of global query systems require a combination of technology, cognition and knowledge, as depicted in Figure 1. The one dimension that has long been neglected is "knowledge," which should be put on-line to provide intelligent assistance to users. We submit that enterprise metadata holds a key for effecting this dimension; that is, knowledge through enterprise metadata is elevated and explicitly formulated to play the central role in a new solution approach to solving the global query problem in this research. Since enterprise metadata are essentially information models, this approach is referred to as the *model-based global query system*. The original concept of such a system as well as its execution methods are the major contributions of this paper.

Specifically, the conceptual model formulates the fundamental needs for enterprise metadata in general terms with which other systems can adopt, with or without the metadatabase. The execution model then avails a new direct method using model traversal to assist end users and a new high-level language using global model constructs to support programmers; both of which are based on a metadatabase enabling the on-line intelligence and assistance capabilities. Automatic derivation of required metadata which the users do not provide is a key element in this approach;

thus new methods such as rule-based consistency checking and message-based search for shortest solution path are developed for the execution model. These results lead to a prototype developed at Rensselaer recently to test the soundness of the approach and demonstrate the global query ability for multiple databases environment.

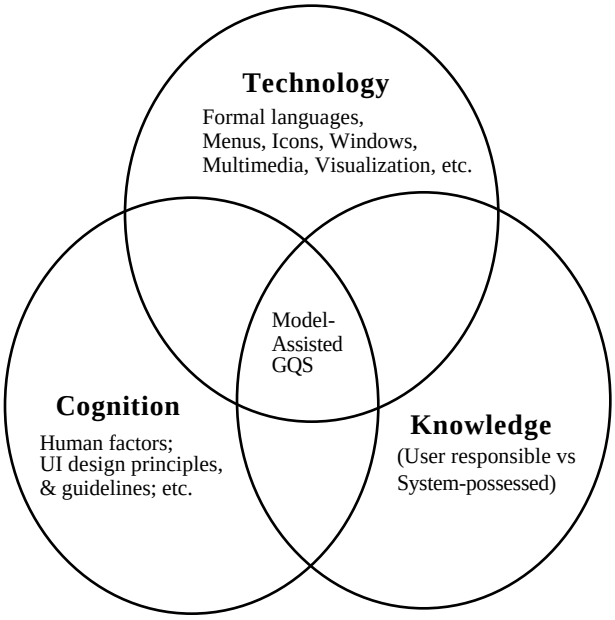


Figure 1 Dimensions of Global Query User Interfaces

The next section presents a conceptual model of this approach, defining the uses and requirements of metadata in a global query system. New methods that are required for the implementation of the approach are provided in Section 3, while the implementation prototype and the new query language using metadata are included in Section 4. The empirical verification of the Model-assisted Global Query System is discussed in Section 5 through using the prototype for an integrated manufacturing case. An analysis of this system vis-a-vis some representative systems (e.g., Multibase, MSQL, CIS, and KIM query system) in the literature is

presented in Section 6. Section 7 concludes the paper with further remarks and future directions of research. Although the prototype system has been reported elsewhere as a part of the metadatabase system for certain industrial applications [29,31], the complete model - both conceptual and execution - and methods are presented and discussed in this paper for the first time. No particular results of the model and methods have been published before.

2 THE MODEL-ASSISTED GLOBAL QUERY APPROACH

The basic logic of the model-assisted global query approach proceeds as follows: First, all classes of enterprise metadata are specified and structured through a metadata representation method. This metadata structure (abstraction) then serves as the basis for organizing and implementing enterprise information models into an on-line and shared metadatabase facilitating all tasks of information integration. As such, the functional views, processes, data models, business rules and other knowledge concerning the global query operation are readily available to both the users and the system through the metadatabase. Therefore, on-line assistance on query formulation and processing becomes a feasible and fully characterized concept. Specific methods based on this knowledge can be defined and developed in terms of metadata requirements and utilized in each major task of the problem.

2.1 The Goals

The target of the Model-assisted (used interchangeable with metadatabase-assisted) Global Query System (MGQS) is delineated in the goals below; which will later be used as the criteria for comparing the MGQS with the existing technologies.

1 Information sharing: achieve information sharing in heterogeneous, distributed, and autonomous environments by means of global queries.

2. Sub-system transparency: support a global model of the whole system and hide local implementations from the users.

3. Local autonomy: maintain local control of its own applications and allow (potentially) for local differences. In addition, it also implies that integration of local systems should not necessitate major conversions or merging of the existing systems.

4. Interoperability: accommodate local heterogeneities while resolving conflicts in data equivalency, different data models and different data manipulation languages.

5 Open system architecture: support the flexibility and adaptability for incorporating new application systems or updating the old ones without imposing major recompilation of reconstruction efforts.

6. Direct query formulation: provide sufficient enterprise metadata to facilitate the articulation and representation of global query using directly the information models via non-command user interface (or minimum-syntax query language for program interface). The user or programmer is not responsible to providing the technical details of the local systems.

7. On-line assistance: use enterprise metadata (including business rules and other contextual knowledge) and knowledge processing capability to assist on difficult tasks for both query formulation and processing. These tasks include, but are not limited to, model traversal and semantic check in direct query formulation, derivation of implied data items and operating rules in query, context-based

joint path optimization, and data equivalence in the assembly of local results.

2.2 The Definitive Model for Metadata Requirements

We first formally characterize the role of metadatabase as on-line knowledge for global Query operation. This characterization starts with a technical analysis of the major global query tasks and their basic metadata requirements. Since these tasks are generic and are not tied to any specific systems, the analysis applies to the general problem studied in this paper.

2.2.1 A global query operation algorithm

Let GQ denotes a global query characterized by a set of attributes/data items (A) that are involved in the query operation; a set of persistent, stored data objects (D) from which all the attributes are drawn; and an expression <C> that specifies the retrieval conditions. Expression <C> consists of sub-expression for selection conditions <SC> and join conditions <JC>. Finally, all data items, objects, and expressions are subject to specification in terms of systems metadata <M>. These metadata, may be either supplied by the user or provided by the global query system, must satisfy a minimum scope required by each particular query. Specifically,

GQ = (A, D, <C> | <M>) where

$A = A^u \quad A^s$ with $A^u \neq \emptyset$,

$D = D^u \quad D^s$ with $D \neq \emptyset$ and $D^u \cap D^s = \emptyset$,

<C> ::= [<SC> | <JC> | <SC> AND <JC>],

<M> ::= {<M^u>} {<M^s>}.

The additional data subsets are explained below:

A^u: This represents the set of data attributes selected by the user. It includes both the items requested directly for retrieval and the items indicated in the selection conditions. A global query must have at least one data item in A^u.

A^s: The system may determine that additional data items are also needed or implied in the query, and hence fill in some data attribute (the set A^s) for the purpose of query processing.

D^u: The user could also specify the set of data objects (D^u) which contain the selected item(s) (i.e., the set A^u).

D^s: Since D^u may not contain all items in A^u, therefore the system will again determine the remaining data objects (D^s) which are involved in the global query operation.

$<M^u>$: This is the user-supplied metadata; which represents the technical knowledge that the user must possess about the enterprise information models and multiple systems in order to represent and sufficiently specify a query.

$<M^s>$: This is the system-supplied enterprise metadata. The set of $<M^s>$ is the complement of $\{<M^u>\}$ with respect to the minimum metadata requirements for a particular query.

The global query operation can be described as a process consisting of the following steps:

Step 1: Global query formulation

Since the user is not necessarily required to specify all of the technical details (the remaining ones will be filled in by the query system automatically), the result of the formulation is likely to be an incomplete global query IGQ defined as:

IGQ = (A^u, [D^u], [<SC>] | [<M>])

where the data objects D^u and selection conditions <SC> may or may be required. For example, a global query: "Find part ID and quantity completed for Jane Doe's order which has a desired date of 5/10/91" could imply the following sets in the formulation step:

A^u = {PARTID, NUM_COMPLETED, CUST_NAME, DATE_DESIRED}

D^u = {WORK_ORDER, CUSTOMER}

<SC> = CUST_NAME="Jane Doe" AND DATE_DESIRED = "5/10/91"

<M> = the exact names and syntax used to specify the elements of A^u, D^u, and <SC>.

A direct approach for end user global query formulation may be employed to formulate the above query via model traversal where the user has the choice of picking as few as only some data items in A^u as many as other information she/he wants to include.

Model traversal

Model traversal is a direct approach whereby users gain the enterprise metadata and utilize them to articulate the query directly in terms of information models. The technical details and semantics of the heterogeneous systems are provided interactively and may be iteratively. The user will, for example, “pick” the data items and objects directly from the models as opposed to “enter” their names to the query. Every step along the way, on-line assistance is provided to the user to traverse as well as pick. Some common semantic errors due to syntax-based translation or interpretation of information models in conventional "indirect" approaches are avoided, as the user sees and deals directly with the comprehensive

and unequivocal system models. The purpose is to allow the user formulate a global query while traversing the information models.

The specific traversal method is designed according to the characteristics of the model constructs and logical associations among the information models stored in the metadatabase.

The following is a basic model traversal process:

Repeat (for each visit)

Step 1.1 Traverse to the data object identified at the i^{th} visit (d_i^u) and select the data item(s) a_i^u (a_i^u from d_i^u) for retrieval.

Metadata required: names of the applications, functional views, data constructs, attributes and their associations (i.e., the global data model).

Step 1.2 Specify selection condition(s) <SC> that will be imposed on a selected data item(s).

Metadata required: formats, domains and operating constraints of the data items

Step 1.3 Resolve semantic ambiguity.

Metadata required: semantic constraints such as functional dependencies and business rules that describe the intended use of the data.

Until no more intended data attributes (i.e., all elements of A^u are specified) $A^u = \bigcup_i (a_i^u)$ and $D^u = \bigcup_i (d_i^u)$.

Step 2: Join conditions determination

A global query may involve multiple data objects that are stored in one or more local systems. Normally, the user has to explicitly specify the equi-join conditions between these data objects for a complete global

query. The specification would require detailed understanding of the information model. To relieve the user of this burden, the system with on-line metadatabase can perform this job through an automatic join condition determination algorithm using enterprise metadata, as follows:

Step 2.1 Determine the set of data objects, O_k, which contain all the user selected attributes, A^u, and their equivalent data items.

This step establishes the maximum space of data objects that the query involves.

metadata required: associations between the entities/relationships and their data items, and data equivalence information.

Step 2.2 Determine a minimum set of data objects, O_{min}, that contain all $a_k^u \in A^u$ and $D^u \subseteq O_{min}$.

Step 2.3 Identify the shortest path (SP_{min}) which connects all data objects $d_m \in O_{min}$.

metadata required: associations between the entities and relationships (global data model).

Step 2.4 Insert join conditions for every connected pair of data objects in D.

The results of step 2 for the earlier example would be:

A^s = {ORDER_ID CUST_ID CUST_ORDER_ID}

D^s = {ORDER}

<JC> = ORDER_ID = CUST_ORDER_ID AND
ORDER.CUST_ID = CUSTOMER.CUST_ID

Step 3: Global query processing

A formulated global query GQ = (A, D, <C>) is decomposed into a set of local queries {LQ_i} where i indicates the local system. The decomposition is

based on the physical whereabouts of the intended data. Each local query LQ_i is pertaining to one and only one local system.

$LQ_i = (LA_i, LF_i, <LG_i>)$ where

LI_i is a set of local items with $i(LA_i) = A$,

LF_i is a set of local files/base relations/record types/objects, and,

$<LG_i>$ is a condition expression concerning only the data item(s) that is contained in the local system i.

Step 3.1 Determine all data files which contain A_j

metadata required: data equivalence, physical storage such as files, relation tables, and their data items (implementation models).

Step 3.2 Determine a minimum set of data files (F_{min}) from which the query system retrieves data items (A).

Step 3.3 Formulate local query LQ_i for local system i, such that

a) (lf_i) $(lf_i \in LF_i)$ $(lf_i \in F_{min})$,

b) (la_i) (lf_i) $(la_i \in LA_i)$ $(lf_i$ contains $la_i)$

c) all items involved in $<LG_i>$ are elements of LI_i.

metadata required: physical storage methods such as files, relation tables, and their data items.

Step 3.4 Preserve global join conditions <GJC> such that

<GJC> ::= <j_condition> [AND < j_condition>]

<j_condition> ::= item1 = item2

where item1 and item2 belong to two different systems.

metadata required: same as Step 3.3.

Suppose, two local systems, shop floor control and order entry system, were involved in the query in the previous example, then two local queries would result from this step:

LQ_{SFC}:

li_{SFC} = {PART_ID NUM_COMPLETED ORDER_ID}

lf_{SFC} = {WORK_ORDER}

<LC_{SFC}> ::= (no condition)

LQ_{OES}:

li_{OES} = {CUST_NAME CUST_ID DATA_DESIRED ORDER_ID}

lf_{OES} = {ORDER CUSTOMER}

<LC_{OES}> ::= CUST_NAME = "Jane Doe" AND
DATE_DESIRED = "5/10/91" AND
ORDER.CUST_ID = CUSTOMER.CUST_ID

<GJC> ::= WORK_ORDER.ORDER_ID = ORDER.ORDER_ID

Step 4: Local query generation

A language generator is needed for each distinctive data manipulation language used in the enterprise. A local query is generated using the local language for each LQ. For the earlier example, the query language LQ_{SFC} generated for the shop floor is in Oracle/SQL:

```
SELECT    WORK_ORDER.ORDER_ID, '|',
          WORK_ORDER.PART_ID, '|',
          WORK_ORDER.NUM_COMPLETED, '|'
FROMWORK_ORDER;
```

The query language LQ_{OES} generated for the shop floor is in Rdb/Rdo:

```
invoke database filename OES$DIR:OES
FOR A IN ORDER
    CROSS B IN CUSTOMER
```

```
            WITH A.DATE_DESIRED =  "5/10/91"
            AND B.CUST_NAME = "JANE DOE"
      PRINT "@",
            A.CUST_ID ,"|",
            A.CUST_ORDER_ID ,"|",
            A.DATE_DESIRED ,"|",
            B.CUST_NAME ,"|"
END_FOR
```

metadata required: implementation models: local DBMS, local DML, and access path, and security metadata: user's access authority and password.

Step 5: Query execution

A local query, LQ must be sent via the network to the destination for processing by the local database, and then theresult will be sent back.

Step 5.1 A message is produced for each local query generated from step 4 containing destination, priority and other metadata, in addition to the LQ.

Step 5.2 These messages are transmitted to the appropriate local system by a network administrator/monitor.

Step 5.3 At the local system shell, message is received by the network administration/monitor and dispatched to the database management system where the local query is executed.

Step 5.4 Local result is sent for Result Integration by the network administration/monitor as a message.

Note: The above description assumes a networking system using the message methods and possessing local as well as global administration

monitoring capabilities. These assumptions are consistent with virtually all network protocols such as TCP/IP, MAP, and TOP.

metadata required: For a minimum network, the metadata requirements can be satisfied by the previous steps. For more advanced systems, metadata such as priority and alternate sources can be used for flows management and optimization.

Step 6: Result integration

The results of local queries (LQ's) must be interpreted and assembled according to the global join conditions (<GJC>) in Step3.4. Logically equivalent data items may be implemented differently in terms of format, scale, and encoding in different local systems. In our example, ORDER_ID in WORK_ORDER and CUST_ORDER_ID in ORDER are encoded differently for local processing purposes. Therefore, data conversions must be performed on one or both of them before the results from these two systems can be joined and presented to the user.

metadata required: data equivalence, contextual knowledge: conversion rules, operation rules.

2.2.2 A minimum model for on-line knowledge

The minimum metadata requirements identified above are organized into a definitive model as follows to characterize the knowledge needed for the global query operation.

Definition: Enterprise metadata gives rise to the knowledge required in global query formulation and processing for both end-users and application programs.

Knowledge for global query formulation

- **Global data model: a logical model representing the data resources of the enterprise. Specifically, the (names of) applications or functions, data constructs and their relationships are needed for model traversal; format and domains of the data items are used for selection condition(s) specification; functional dependencies are for integrity checking of the selection conditions, and primary keys, and foreign keys are for implicit join conditions determination.**
- **Data equivalence: the knowledge needed in most steps to convert as well as identity/clarify multiple data definitions; including typing, semantic (interpretation of) presentation, and scale.**
- **Contextual knowledge: business rules describing the intended use of the data and the needs of the user, also used for ambiguity checking of the selection conditions.**

Knowledge for global query optimization and decomposition

- **Data equivalence.**
- **Implementation models: physical storage methods of the logical data items, including the size of the files or relation tables and the like used for query optimization and decomposition.**

Knowledge for global query generation

- **Implementation models: metadata about local data language environments and access paths, used to determine the language generators to use and the heading of a query program.**
- **Security metadata: users' access privileges, used to determine whether the retrieval requests are legitimate and the passwords are used for obtaining access permission.**

Knowledge for results integration

- **Data equivalence**

- **Contextual knowledge: conversion rules and operating rules needed for resolving conflicting data definitions and computing derived data items.**[1]

2.3 The Approach: The Conceptual Model

The above analysis, while applies to the global query problem in general, also defines the overall algorithm and the minimum contents of the metadatabase for the model-assisted global query approach. Thus, a defining characteristic of the MGQS approach is: S{} ⊃ M{<M>}; i.e., the system provides all of the metadata required.

The methods that are required to implement this approach are discussed next.

3. NEW METHODS: UTILIZING THE METADATABASE FOR ON-LINE ASSISTANCE

The execution methods envisioned in the model-assisted GQS are developed below. Their objectives are, as discussed before, utilizing a metadatabase to provide on-line intelligence and assistance to facilitate global query formulation and processing. The metadatabase can be designed in a number of ways, of course. The particular execution model we developed is based on the particular Metadatabase developed in the past several years, whose structure is detailed in [28]. The metadatabase structure itself is outlined first.

3.1 The metadatabase Structure: the GIRD Model

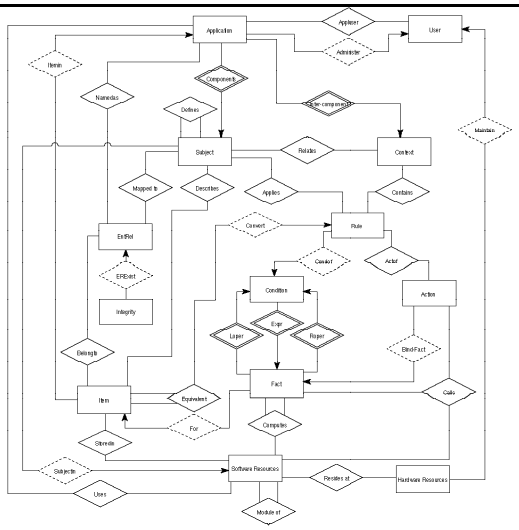


Figure 2 The GIRD Meta-Model

The metadatabase itself employs a generic meta-structure, the Global Information Resources Dictionary (GIRD) model (see Figure 2), abstracting the enterprise metadata resources (i.e., models) into a metadatabase. As such, each and every local model is represented (but not duplicated, nor removed) into the metadatabase as ordinary metadata "tuples" populating the structure. Therefore, metadata independence [29] at the model integration level is achieved since any change in enterprise models would involve only ordinary metadata transactions similar to the usual relational processing, and do not require any change to the structure itself, nor reloading/recompilation. The term metadata independence is phrased in the same philosophy as the well-known data independence concept. The structure itself, as well as the global representation of local models as metadata instances, is developed using the Two-Stage Entity Relationship (TSER) method, which is reported elsewhere [24, 25, 30]. TSER provides at the first, functional stage two basic representation constructs: SUBJECT (similar to objects in objects in object-oriented paradigm) and CONTEXT

(rule-based description of process and other contextual knowledge). At [3] the second, structural stage, ENTITY (characterized with singular keys) and RELATIONSHIP (characterized with composite keys) plus two more types of associations signifying special integrity rules (functional and mandatory) are defined. All these constructs, combined with hardware and software classes of metadata, are included in the GIRD model. (The above description suffice the purpose of this paper; the full details are provided in [28].)

This design provides a few basic properties important to the execution model.

(1) The GIRD model can be implemented in the same manner as a regular schema using a relational DBMS, where the TSER constructs of the model provide design specifications.

(2) Local information models (represented in the metadatabase as tuples) can be added, deleted, or modified without causing the metadatabase to restructure nor recompile.

(3) All meta-relations in the GIRD model (e.g., APPLICATIONS, SUBJECT, DEFINE, CONTEXT, RULE, ITEM, EQUIVALENT, and BELONGTO) are normalized. Thus, they can be managed and processed as a (relational) database implementing the model.

(4) Contextual knowledge is represented in terms of relations as well. A particular class is the equivalence between data items. For instance, the fact that 5/31/94 in the American format of date is equivalent to the European 31/5/94 is established through ITEM and EQUIVALENT, with the attendant conversion rules and routines represented through RULE.

(5) The required metadata as defined in section 2 are all included in the GIRD model. Thus, Figure 2 shows the high-level semantics of the metadatabase.

(6) In addition to effect a full-fledge metadata management facility, the metadatabase also simplifies the mapping requirements. Since all local databases map directly with the metadatabase rather than among themselves, the complexity is N as opposed to N^2.

(7) The modeling and creation of metadatabase follow exactly the tradition of data and knowledge systems analysis and design. Full discussions of a particular methodology and its CASE implementation can be found in [29, 30].

(8) The metadata independence nature of the architecture supports naturally scalability in terms of adding new systems; while its implementation using a regular DBMS assures scaling up in software engineering. This property is also discussed in [29,30].

3.2 Query Formulation

3.2.1 Model Traversal

Two basic methods are employed for model traversal: vertical and horizontal (Figure 3). The vertical method specifies the traversal depth cutting across application systems, functional models, structural models, and data items. The horizontal method, on the other hand, traverses the global model from an entity/relationship (ER) to others in a network manner.

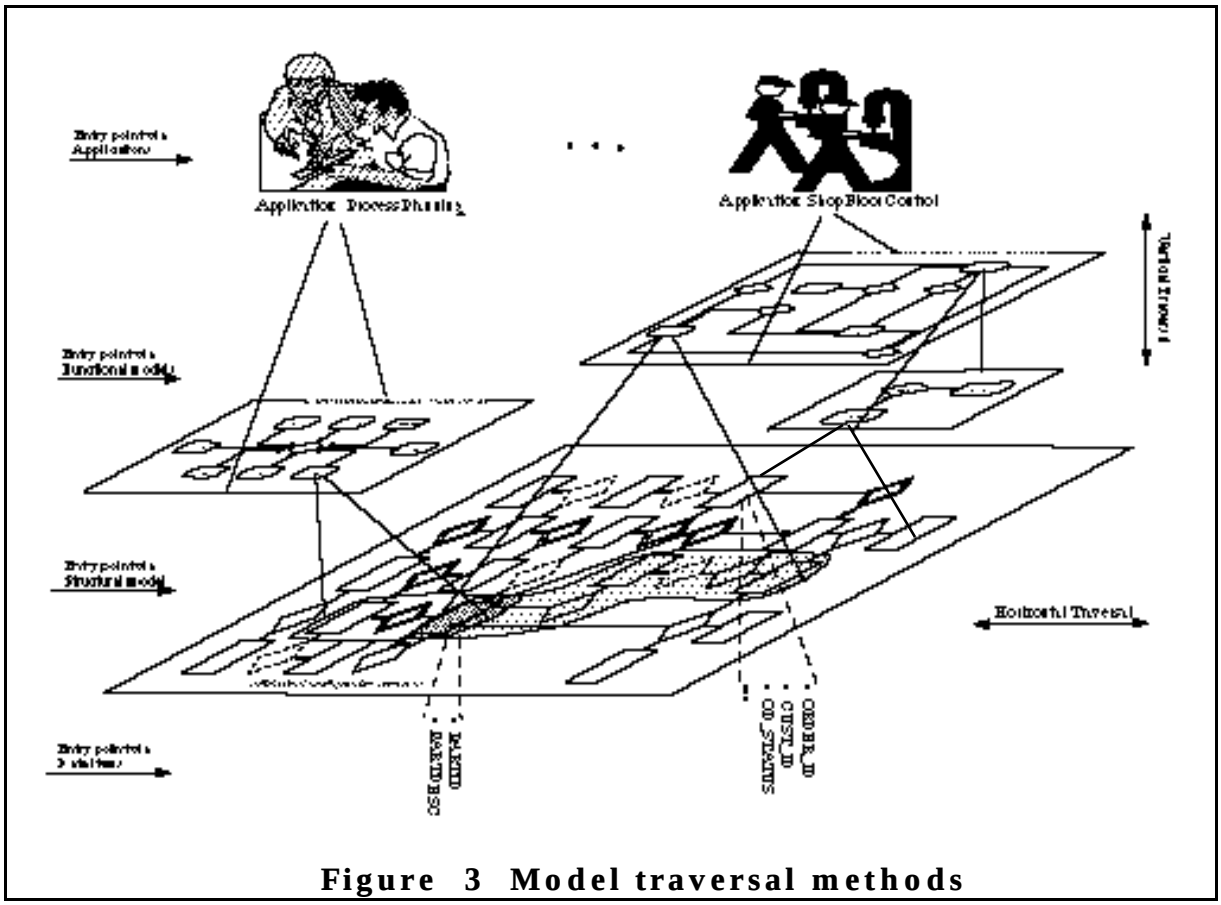


Figure 3 Model traversal methods

The user could page through the global model according to the logical associations between these modeling constructs. Depending on the user's experience with the information models, the traversal could be very pinpointing (specifying the exact constructs containing the data items needed) or very general (browsing through a list of data items with little precise specification or through only the applications), or anything in between.

Selecting E/R and data items

The result of a simple selection in global query formulation is equivalent to the projection operation of relational algebra which reduces the number of columns in a table. For instance, selecting data items a and b from ER1 can be expressed as:

ER1 projected_to a, b{<r.a,r.b>|ER1(r) }

where r is a row (i.e., tuple) of an ER.

If the result of the selection involves data items contained in multiple ERs, the interpretation would be different depending on whether or not the automatic equi-join determination is used. The automatic determination will leads to first performing equi-joins among the ERs and then the projection on the results of the joins. For example, consider a case where data items a and b are selected from ER1, and c and d are selected from ER2. Assuming ER1 and ER2 are directly connected in the global model (i.e., they share a common item (x) as part of their primary key or foreign key), the operation can be expressed as:

1. ER1 join ER2 on x: ER = {r++t-<t.x>|ER1(r)ER2(t) (r.x=t.x)}

   r++t denotes a row made by concatenating row t onto the end of row r

2. ER projected_to a, b, c, d: {<u.a, u.b, u.c, u.d>|ER(u)}.

If the selected ERs are not directly connected to each other, then a connected solution path needs to be determined (see Section 3.2.1) in order to determine the necessary equi-joins.

When the equi-join determination is not used, it implies a Cartesian product (*) between ER1 and ER2 is performed among the ERs to precede the projection operation. The same example would be expressed as:

1. ER1 * ER2:ER = {r++t|ER1(r) * ER2(t)}
2. ER projected_to a, b, c, d: {<u.a, u.b, u.c, u.d>|ER(u)}.

Selecting derived data items

Besides selecting the persistent (stored) data items, the user can also include derived (run-time) data items in a global query. A derived data

item is not stored in any local system, but computed by using the persistent data item(s). Continue the above example, data item x = f(a,c). The derived item (x) can be selected from either ER1 (ER1 projected_to x) or ER2 (ER2 projected_to x). The result of the selection causes both a and c to be included in the global query:

{<t.a>|ER1(t)}, {<r.c>|ER2(r)}.

The derived data items are defined via business rules in the metadatabase. The computation of derived data items using a rule processor is discussed in Section 3.5.

Specifying selection conditions

Each selection condition (f(a)) imposes a restriction on an data item a. For example, specifying the selection conditions "f(a) (g(b) h(c))" for an entity/relationship ER1 containing a, b, c would mean:

{r|ER1(r) f(r.a) (g(r.b) h(r.c)) }.

If, on the other hand, c is contained in ER2, then there are two different actions depending on whether equi-join determination is used. If the equi-join determination is used, it means to first perform an equi-join between ER1 and ER2 on the common item x:

ER = {r++t-<t.x>|ER1(r)ER2(t) (r.x = t.x)}

and then a selection on ER:

{u|ER(u) f(u.a) (g(u.b) h(u.c)) }.

If the equi-join determination is not used, it means to perform a Cartesian product between ER1 and ER2:

ER = {r++t|ER1(r) * ER2(t)}

before the selection operation.

A successfully formulated global query can be kept as a run-time view by giving a unique name, which can then be recalled to formulate a nested global query. MGQS supports two different methods of relating run-time views: UNION (+) and NOT EXISTS (-). The user starts with selecting a method and two views to combine. Then the user selects items from the set of common items of these views presented by the system. The result of the UNION of views V1, V2 based on data items a, b (V1 + V2 on a, b,) could be expressed as :

$\{r++t - \langle \text{duplicate columns} \rangle | V1(r)_{a,b} V2(t)\}$.

The result of V1 NOT EXISTS (i.e., NOT EXISTS works as the relational algebraic operator DIFFERENCE) in V2 based on data items a, b (V1 - V2 on a, b,) could be expressed as:

$\{r|V1(r)_{a,b} \sim V2(t)\}$.

3.2.2 Query Validation

MGQS detects and prevents inconsistencies in the selection of join conditions by examining the knowledge in metadatabase. Specifically, assistance is provided for the following three potential problems:

(1) Domain incompatible conditions: The system will use the properties of data domains to guide user's entry of values for a selection condition. It will only allow values that conform with the type(s) (real, integer, character, etc.) of the data item(s) selected.

(2) Semantically inconsistent conditions: The formal semantic constraints such as functional dependencies model are used to detect and prevent semantic inconsistencies.

(3) Conditions conflicting business rules. Business rules might implicitly nullify or render certain selection conditions that the user intends invalidate. The system will perform this consistency checking via the rule processing method provided in Section 3.4.

3.3 Automatic Completion of Global Query Formulation

The MGQS method determines all of the required metadata that users do not provide to complete the query after model traversal. A key task here is the automatic determination of the join conditions implied in user's (incomplete) formulation of query by using the global data model from the metadatabase. The data items and ERs that are not originally selected for the global query, but that are needed for the joins will be added to the global query for completion.

The steps of metadata derivation for completing a global query include :

3.3.1 Determining a minimal set of data objects

Equivalent(a): This is a function that requests the metadatabase to retrieve all equivalent data items of the specified data item, a.

DataObject(A, subject, application): This is a function that requests the metadatabase to retrieve an exclusive set of data objects within the boundary of the specified subject and/or application, such that every resulting data object contains at least one data item in A.

The minimum set of ERs algorithm

Step1. Set $O_{min} = D^u$;

Step2. For (each $a^u_k \in A^u$)

O_k = DataObject(Equivalent(a^u_k), subject, application);

Step3. For ((each $o \in O_{min}$) and ($A^u \neq \emptyset$))

For ((each $a^u_k \in A^u$) and ($A^u \neq \emptyset$))

```
                if o   Ok
                        Au = Au - {auk };
Step4. if Au ≠    after Step 3 then Step 4.1 and Step 4.2
   Step4.1 For ((aul    Au) and (||Ok|| = 1))
                {
                Au = Au - {auk };
                Omin = Omin  Ok;
                For each aul    Au and Au ≠
                        If (ok   Ol)
                                Au = Au - {aul };
                }
   Step4.2 get_minimal_sets (Omin, Au, results)  /* results = {} initially  */
        {
        current_best = ||Au|| + ||Omin||
        select auk  from Au;
        for ((each ok   Ok) and (Ok ≠  ))
                {
                If Au - {auk } =
                        If ||Omin   {ok}|| < current_best
                                {
                                results = {Omin   {ok}};
                                current_best = ||Omin   {ok}||;
                                }
                        else If ||Omin   {ok}|| = current_best
                                results = results {Omin   {ok}};
                else
                        get_minimal_sets(Omin   {ok}, Au-{auk }, results);
                }
        }
```

3.3.2 Shortest Solution Path Determination

In the context of global query formulation, the global data model is considered as a network G which is a graph where all nodes (the set N) are connected with links (the set L). The set of selected entities and relationships representing the semantic of a global query defines the set of nodes (N') to be connected by the solution path sought. The shortest solution path (SG) which connects all selected entities and relationships

denotes the complete formulation of the global query. An algorithm is presented for searching for the SG. Although there exists generic algorithms and even specialized entity-relationship algorithms for the SG problem, a new idea is employed to develop a, hopefully, more efficient SG algorithm to deal with large networks.

The algorithm begins by creating messages; one message for each of the node in N. A message contains a unique identifier of the message (ID), the cost (cost), and an indicator (from) signifying the preceding node of the message. When the name of the node (i.e., entity or relationship) in the global data model is unique, this name can be used as the ID; otherwise, the system would generate unique identifiers. The cost of a message represents the number of links that a message has been passed through, from the originating node (i.e., one in N) to the current node. Lastly, "from" contains the ID of the previous node that sent out the message.

The algorithm centers around messages. The idea is to send a message from the originating node to all its adjacent nodes, which, in turn, pass each message to new nodes one unit distance farther from their originating node. While two messages with same IDs reach a node, only the one with smallest cost is kept. This cycle continues until a node has collected all messages that were originally created; at this point a solution path has been found. The total cost of the path is then calculated by adding all costs of all collected messages. If the total cost of a new solution is less than the cost of the previous solution, a better solution is found. The new total cost is then recorded and the node is marked as the current root R of the solution. This solution searching process is continued until the shortest path is found. Two criteria are used to terminate the algorithm: (1) current_cost = nb_nodes -1 (current total cost equals to the

number of nodes in N minus one) which means all nodes are directly connected, (2) nb_cycle = current_cost - nb_nodes + 2, where nb_cycle is the number of cycles for message sending. The formal algorithm is given below, which is followed by a proof for the termination conditions.

Lemma 3.1. A lower bound for the best solution (shortest path) that can be found for connecting the nodes in N has total cost C_R = $||N^o|| - 1$, where $||N||$ is number of elements in N.

Proof. The best solution occurs only when the root node R is in N^o and all other nodes k ($k \neq R$, $k \in N^o$) are directly connected to R.

Thus, $l_R = 0$, (l is the distance from a node to the root)

$$l_k = 1, \quad k \in N^o, \text{ and } k \neq R$$

$$C_R = \sum l_n = \sum l_k + l_R = ||N^o|| - 1 \blacklozenge$$

Lemma 3.2. The longest possible distance (L^m) between a farthest node n_1 ($n_1 \in N^o$) to the root node R in a solution with the total cost of C_R is $L^m = C_R - ||N^o|| + 2$.

Proof. Let n_1 be the farthest node from the root R, and $K^o = \{n_1\}$ (Figure 4).

$$C_R = \sum l_n = \sum l_k + l_{n1}, \quad \text{where } \forall k \in K, \text{ and } k \neq n_1$$

$$\Rightarrow \quad l_{n1} = C_R - \sum l_k$$

From Lemma 3.1, the smallest possible cost to connect k nodes is $||K||-1$, where $K = N-1$. Thus $L^m = MAX(l_{n1}) = C_R - ||N^o|| + 2 \blacklozenge$

Theorem 3.1. The algorithm will reach the shortest path (all shortest paths if more than one exist) at the end of i message sending cycle, where $i = C_R - ||N^o|| + 2$.

Proof. From Lemma 3.2, the largest distance possible between a node (n_1) to the root (R) is $l_{n1} = C_R - ||N^o|| + 2$. Each cycle of message sending will pass the message (m_{n1}) one distance closer to the root. Thus exactly i (i =

l_{n1}) cycles are needed for the farthest message m_i to reach R. Also, if a better solution with root R' exists ($R' \in C_R$), the number of cycles required to reach the solution i' = $i - ||N^o|| + 2$ => i' < i.

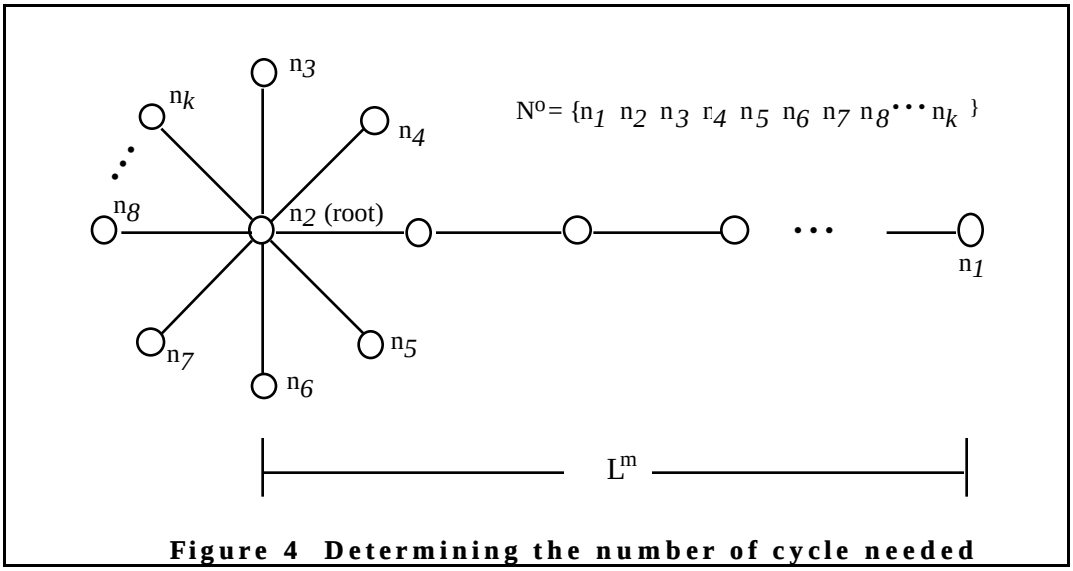


Figure 4 Determining the number of cycle needed

The shortest path algorithm

```
{
current_cost = a large number;
root = NULL;
create message for each node in N^o;
initialize current_cycle;
next_cycle = NULL;
initialize visited_nodes;
nb_cycle = 1;
for ((current_cost > nb_nodes -1) and
      (nb_cycle ≤ current_cost - nb_nodes + 2))
      {
      for (each message m_i in current_cycle)
            {
            list_of_nodes = get_related_entries(n_from_i);
            for (each node n_j in list_of_nodes)
                  {
                  if (n_j  visited_nodes)
                        {
                        add n_j to visited_nodes;
                        m_i ->cost = m_i -> cost + 1;
                        m_i ->from = n_j ->key;
                        add m_i to next_cycle;
                        }
```

```
                    if (m_i not in n_j)
                        {
                        add message m_i to n_j;
                        if (new_best_solution())
                            {
                            root = n_j;
                            current_cost = calculated_cost();
                            }
                        }
                    }
                }
        current_cycle = next_cycle;
        next_cycle = NULL;
        nb_cycle = nb_cycle + 1;
        }
```

Semantic Ambiguity on Solution Paths

Different solution paths represent different semantics. When the shortest path determination algorithm returns two or more shortest solution paths, it indicates further interaction with the user is needed to clarify the ambiguity. A straight forward method is to display all the alternative paths (graphically)with suggestions, interpretations, and possibly also dialogues. Then, the user will select the correct path from the alternatives. Towards further automation, the MGQS will match the ERs involved in a path with the ERs of the SUBJECTs and user defined views stored in the metadatabase. If a match is found, the matching path is likely to be the correct one. Therefore, a recommendation could be made to the user. The idea is to offer more assistance by providing, or acquiring, more information about each path and the matching with SUBJECTs and views. However, the user will have to make the final decision for the correct solution path.

3.3.3 Equi-join Conditions Determination

To complete the query, join conditions needed to implement the solution path found must also be determined. The method starts from the root node. It backtracks the message passing route (i.e., following the value of "from") for each message collected in the root to its originating node (i.e., when from = NULL). Along the route, join conditions are determined for each pair of adjacent nodes. Two routes may share a common section. Therefore, a list of processed adjacent pairs is maintained and the results from the previous join determination process between the same adjacent nodes is used to improve performance.

The specific join condition between two nodes (ER1 and ER2) is determined based on the type of connection between them this is model structure-specific. A join condition is added to the GQ when matching items (i.e., two equivalent items) are found from the two nodes. The particular methods for the TSER model used in the MGQS prototype (see section 4) can be found in [12].

3.4 Query Processing

3.4.1 Query Optimization & Decomposition

Two heuristics are used for global query optimization: (1) minimizing the number of equi-joins between ERs, and (2) minimizing the number of local files being accessed. During the query formulation, data items may be selected from SUBJECTs, applications, or the enterprise as a whole instead of explicitly from ERs. Determining a minimal set of ERs that contain all of these data items and provide the shortest solution path will minimize equi-joins. Determining a minimal set of files is done in the following way:

Step 1. Determine the set of files (F_k) containing the data item (a_k).

For each $a_k \in A$, the set of data items selected by the user and determined for the join conditions by the system:

F_k = GET_FILE(a_k);

GET_FILE() returns all files containing a_k from the metadatabase.

Step 2. Determine the minimal set of files that contain all A and A^u.

Same as step 4 in the minimal set of ERs algorithm, but replace ERs with files, O_{min} with F_{min}, O_k with F_k, a_k^u with a_k, and A^u with A.

Step 3. Determine the set of data items (A^f) that will be retrieved from each file, ($f \in F_{min}$).

Decompose global query into local queries

The purpose of the decomposition procedure is to break down the global query into sub-queries, such that each sub-query concerns only one local system. The decomposition procedure has the following steps:

Step 1. Decompose the global query (GQ) into multiple global queries (GQ_i). Let $A^F = \bigcup_{f \in F_{min}} \{A^f\}$, $f \in F_{min}$.

Let GQ = {F_{min}, A^F, $C(A^c)$}, transform the logical expression $C(A^c)$ into disjunctive normal form using the laws of Boolean algebra:

$C(A^c) = C_1(A^c) \vee C_2(A^c) \vee \ldots \vee C_i(A^c)$

where $C_i(A^c)$ is an elementary conjunct of the form $X_1 \wedge X_2 \wedge \ldots \wedge X_i$.

The result of this step is GQ_i = {F_{min}, A^F, $C_i(A^c)$} (i.e., $GQ = \bigcup_i GQ_i$).

Step 2. Decompose each GQ_i into local queries.

Step 2.1 Group file $f \in F_{min}$ by application (local system)

```
For each f ∈ Fmin
  {
  s = GET_APPL(f);  /*returns the application that file f belongs to
*/
  Fs = Fs ∪ {f};
```

}

Step 2.2 Separate conditions $J_i(A^c)$ by applications.

Let $C_{si}(A^c)$ be the condition expression that concerns only application system s.

$$C_i(A^c) = (\bigwedge_{si} C_{si}) \wedge J_i(A^c)$$

where $J_i(A^c)$ are join conditions that involve item from two different applications.

The result of this step is a set of local queries (i.e., $LQ_{si} = \{Q_s, A^F, C_{si}(A^c),\}$) and the join conditions $J_i(A^c)$.

3.4.2 Generating Local Queries

A local query using the local data manipulation language (or local retrieval language for file system) is generated for each $LQ_{si} = \{Q_s, A^F, C_{si}(A^c),\}$. A code generator is designed specifically for a particular DML (e.g., SQL, Rdo/Rdb, and dBASE languages) based on the structure of LQ and the syntax of the DML.

3.4.3 Result Integration

Step 1. Assemble local results R_{si} into the result of global query GQ_i

$R_i = \{r1++r2++... | R_{1i}(r1) \wedge R_{2i}(r2) \wedge \ldots \wedge J_i(A^c)\}$

r1 and r2 are the rows in local results $R_{1i}(r1)$ and $R_{2i}(r2)$ respectively, and r1++r2 denotes a row made by concatenating row r2 onto the end of row r1.

Data conversion is needed during these join operations, if the two items involved in a join condition are equivalent data (Section 3.4).

Step 2. final result, $GQ_i = \{r | R_1(r) \wedge R_2(r) \wedge \ldots\}$

Derived data items are computed according to the business rules, at the end of this step. The persistent items used for the derivation will be replaced by the derived data item (Section 3.5).

If the global query GQ is formulated by relating two run-time views using the operators UNION or NOT EXISTS, the result would be:

1. UNION views V1, V2 based on data items d.

   $R_{GQ} = \{u|V1(u) \ _d \ V2(u)\}$

2. V1 NOT EXISTS in V2 based on data items d.

   $R_{GQ} = \{u|V1(u) \ _d \sim V2(u)\}$

3.5 Knowledge Processing Using a Rule-based Approach

3.5.1 Convert equivalent data

Function Join_condition_satisfied() returns TRUE if the two rows in the join operation satisfy all conditions in $A^c_{ability}$, and returns FALSE otherwise.

```
Join_conditions_satisfied(rowA, rowB, A^c_ability)
      {
      For (each condition (itemA operator itemB) in A^c_ability)
            {
            valueA = extract(rowA, itemA); /* returns value of itemA in rowA
*/
            valueB = extract(rowB, itemB);
            C_rules = find_convert_rules(itemA, itemB);
            If (C_rules ≠ NULL)
                  valueA = convert_item(itemA, itemB, valueA, C_rules);
            If (compare(valueA, valueB, operator) == FALSE)
                  return(FALSE);
            }
      return(TRUE);
}
```

The function find_convert_rules(itemA, itemB) searches the equivalent table of the metadatabase for the conversion rule(s) (Figure 5) that could convert the value of itemA (valueA) into the format of itemB. As an example in Figure 6 find_convert_rules(itemA, itemB) returns two rules (i.e., rule2 and rule4), which will first convert the value of itemA to itemC format (using rule2) then to itemB format (using rule4). The function returns "NULL" if no conversion is needed for comparing itemA and itemB.

The function covnert_item() triggers the rule processor which will search and fire the actions of the conversion rules [9].

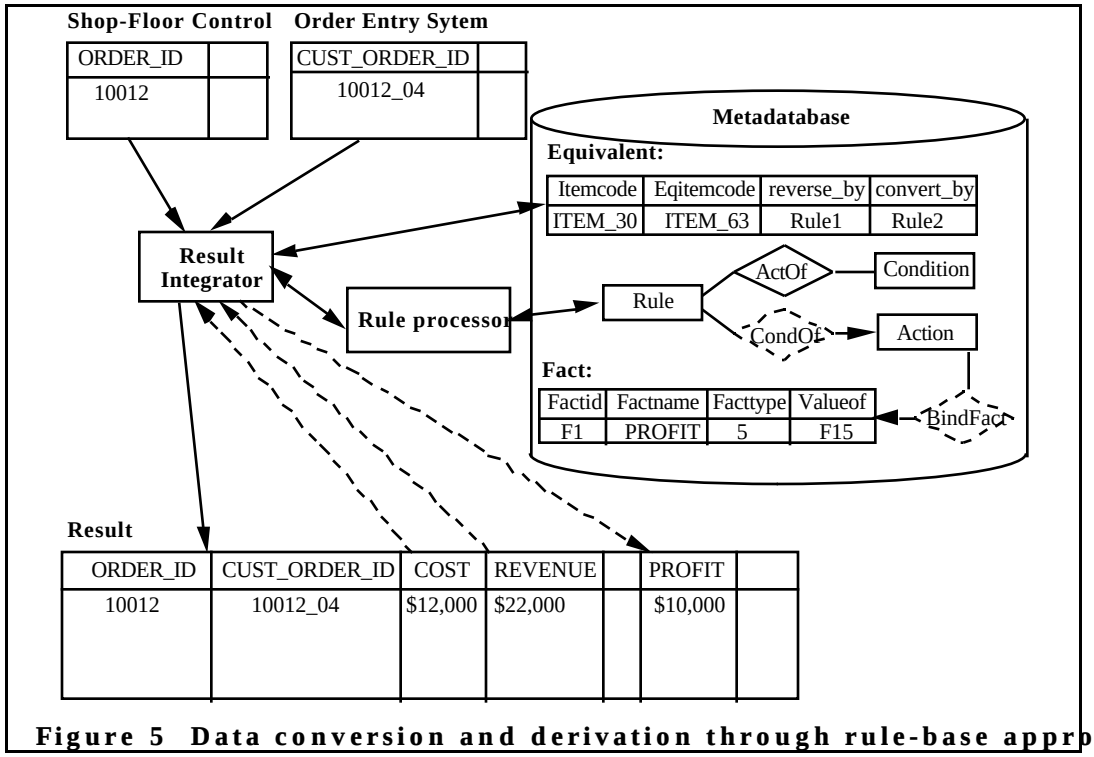


Figure 5 Data conversion and derivation through rule-base approach

Equivalent table

Itemcode	Eqitemcode	reverse_by	convert_by
itemA	itemC	Rule1	Rule2
itemB	itemC	Rule4	Rule3

Figure 6 Equivalent table

3.5.2 Compute derived data item

Function get_involved_items(d, D_p) returns the list of persistent items D_p and the action ID (Actid) of the action that binds a value to d. For example, as shown in Figure 4, the search starts from identifying a fact (Factid) that binds a value to PROFIT (i.e., factname = "PROFIT" and Facttype = 5) in the Fact table of the metadatabase. Using the Factid through Action table we find all the rules (R) that may cause the binding. Therefore, all the involved persistent items D_p can be identified through the Actions and Conditions that are related to R (see Chapter 9 in [9]). D_p is first used for global query formulation and retrieval then the values of D_p (V_{dp}) are used to compute d.

Let Result be the result from assembling the local results and V_{dd} the computed result of the derived item.

```
Compute_derived_item()
    {
    For each row (r) in Result
        {
        ActionID = get_involved_items(d, Dp);
        Vdp = extract(r, Dp);
        populate_fact_table(Dp, Vdp);
        Vdd = BChainer(ActionID);
        insert (row, Vdd);
        }
    }
```

The procedure populate_fact_table() builds the fact table using the persistent items code and their values, V_{dp}. Function BChainer(ActionID) triggers the rule processor which uses the backward chaining to try to reach the goal (ActionID) according to the fact table. The computed value of the derived item is returned by the function.

3.5.3 Check business rules

Consider the following example. Rule12 and Rule14 are stored in the metadatabase;

Rule12:

If WO_SEQ.WS_ID = '0001'

Then WORK_ORDER isa Milling job

Rule14:

If WORK_ORDER isa Milling job And

WO_SEQ.START_DATE ≠ WO_SEQ.END_DATE

Then PRINT("Warning: END_DATE should be the same as START_DATE for a milling job")

If the users select a condition such as:

WO_SEQ.WS_ID = '0001' AND

WO_SEQ.START_DATE ≠ WO_SEQ.END_DATE

then MGQS will detect the inconsistency and issue a warning to the users. This method calls for a forward chaining rule inferencing strategy for the rule processor. The consistency checking can be triggered whenever a new condition is specified or by request (e.g., when the user request to processing the global query).

Step 1: Populate the fact table.

Step 2 Activate the forward chainer.

The function FChainer() triggers the forward chainer which will fire rule(s) according to the fact table.

Note: details of all of these functions described in this section can be found in [9,12]

4 THE IMPLEMENTATION

4.1 The MGQS Architecture and Prototype Method

The general approach discussed above is implemented with particular designs. The metadatabase and its attendant concurrent architecture [4, 26, 27, 29, 31] Provide the technological basis for the particular MGQS methods.

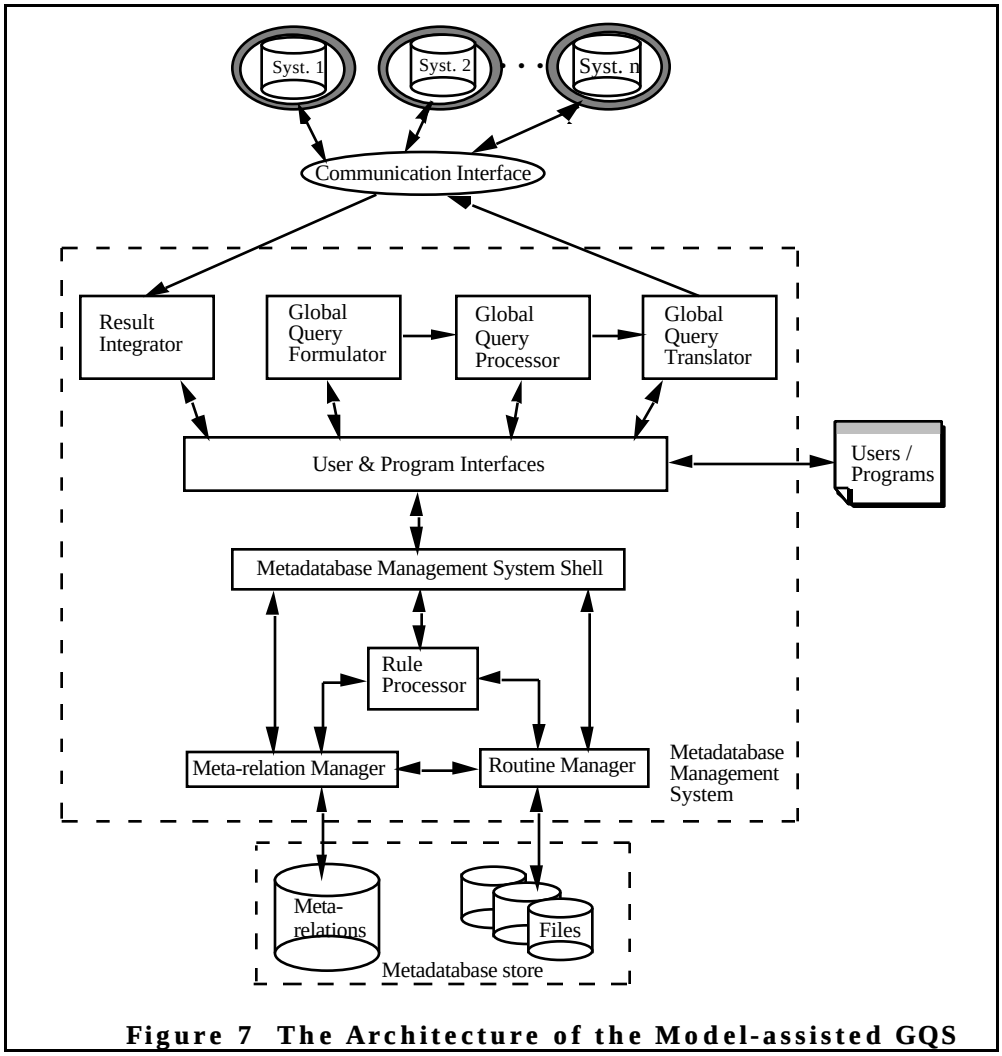


Figure 7 The Architecture of the Model-assisted GQS

An MGQS architecture which complies with the basic requirements described in the earlier sections and the idea of model assistance is developed as shown in Figure 7. It can be conceptually divided into two major components (1) metadata manager and (2) global query manager.

The metadata manager consists of the metadatabase management system shell, rule processor, meta-relation manager, and routine manager [9]. It acquire pertinent knowledge from the metadatabase, and thereby provides model assistance for the global query operation through the user interface. The global query manager on the other hand, includes the necessary modules, such as query formulator, processor, translator and result integrator that would fulfill the basic requirements of global query operations.

4.2 Metadatabase Query Language

The Metadatabase Query Language (MQL) allows programs to query a collection of autonomous local information system(s) in a non-procedural way. Its ability, supported by the metadatabase, to accommodate the logical and physical heterogeneities of the local systems distinguishes MQL from other distributed database query languages.

The major functionalities which MQL provides are the following:

- retrieve metadata against the metadatabase,
- support queries requiring joins of data from different local systems operating under different schemata,
- support queries involving data with multiple definitions (i.e., differing names, formats, units, and/or value coding) within a single system or across different local systems,
- use names with which users are familiar in queries to qualify data elements in different databases,
- minimize technical details expected of users for the global query formulation while supporting the above functionalities (e.g., the physical locations, local names, and implicit join conditions).

The conceptual structure of MQL is based on the Two-Stage Entity-Relationship representation method. In Figure 9, the GET command is the only command which must be stated in a global query. It is used to specify the list of items for retrieval. d_i denotes the list of item names which are delimited by a blank character "space" and NULL is a symbol to indicate that no data item will be retrieved. In the syntax expression, v_i denotes a run-time view name, a_i denotes an application name, s_i denotes a subject name, e_i denotes an entity name, and p_i denotes a plural relationship name.

The FROM clause is used to indicate where the items are to be retrieved from. There are four ways of using the FROM command; FROM VIEW, FROM APPLICATION, FROM SUBJECT, and FROM E/R, depending on the level of detail a user specifies in the global query. If the items are selected from a user defined run-time view, the view name must be provided by using the FROM VIEW clause. Otherwise, FROM is optional in this syntax, especially when the user does not know where specific data items can be retrieved from. During the global query formulation, the FROM and GET commands can be used repeatedly until all of the items to be retrieved are specified (see Figure 10 for an example).

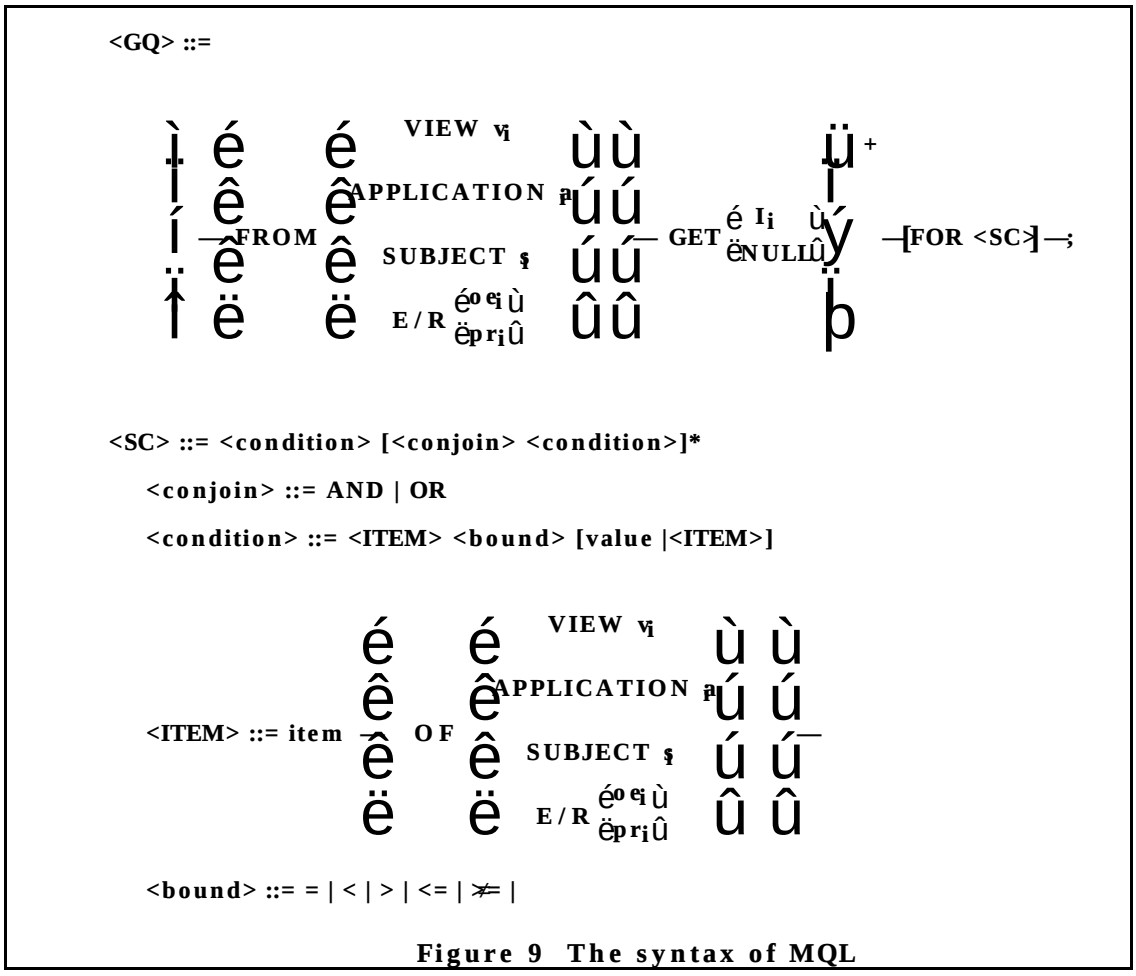


Figure 9 The syntax of MQL

Lastly, the FOR clause is used for specifying the selection condition(s) and join condition(s) (<SC>). The conditions are conjoined by the logical operators AND or OR. Each condition states the binding of an item with a constant value or another item. The logical location of an item involved in a condition can be further specified with the OF clause. The OF clause works in the same manner as the FROM clause.

```
  Global query:      "Find part ID and quantity completed for Jane
Doe's order which
              has a desired date of 5/10/91"

 FROM APPLICATION    SHOP_FLOOR GET PARTID NUM_COMPLETED
 FROM SUBJECT   CUST_ORDER GET DATE_DESIRED
 FROM OE/PR   CUSTOMER GET CUST_NAME
 FOR   CUST_NAME = "Jane Doe" AND
       DATE_DESIRED = "5/10/91";
```

Figure 10 An MQL example

Defining Run-time View

A run-time view is a user defined information view based on the existing information model. After its creation, the view is kept in the main memory as a virtual relational table during the global query operation. To define a run-time view, the user uses the command

DEFINE VIEW v_i <GQ>;

where <GQ> is a global query in MQL.

v_i is the name given to the new defined view. In essence, v_i is the result of a global query.

A global query can be formulated based on a run-time view (using the FROM VIEW and OF VIEW clauses) or by relating run-time views using the UNION and EXIST functions (described below). With the MQL syntax, a view can be built recursively on another run-time view. The capability of view definition allows complex (nested) global queries to be formulated using the MQL.

Standard Functions

Standard functions are provided in MQL, including

AVG() - average of the values (numeric values only) in the column,

COUNT() - number of values in the column,

MAX() - largest value in the column,

MIN ()- smallest value in the column, and

SUM() - sum of the numeric values in the column,.

Three functions DISTINCT, GROUP BY, and ORDER BY are provided for arranging the resulting tables:

DISTINCT - eliminates tuples with duplicate values in the column

GROUP BY - arranges tuples into groups such that within any one group all tuples have the same values for the column

ORDER BY - sorts tuples (ascending or descending) according to the values of a column.

Lastly, UNION and NOT EXISTS are used in MQL to formulate nested global query by relating two run-time views. The syntax of UNION and NOT EXISTS are as follows:

<view$_i$> UNION <view$_j$> on I_i

<view$_i$> NOT EXISTS <view$_j$> on I_i

I_i is a list of data item(s) contained in both view$_i$ and view$_j$. UNION works as the union operator of set theory; the result of a UNION operation is a table with all unique data items from both views. Rows with duplicate values on I_i are eliminated from the result of the UNION. NOT EXISTS, on the other hand works as the difference operator of set theory. The result of a NOT EXISTS operation is a table containing rows from view$_i$ such that there is no value of its I_i that matches any values of I_i in view$_j$.

4.3 Implementation of the Design

A prototype model-assisted global query system based on the architecture described in the previous section is built for the purpose of concept verification [4, 9, 12, 41]. This prototype is written in the C

language and has two versions running respectively on a Micro VAX workstation and an RS6000 workstation. It provides functions of global query formulation, global query optimization, local query translation, result integration, and on-line intelligence using rule-based approach.

The prototype MGQS user interface is designed according to the model traversal method described in Figure 3. Figure 8 illustrate the implementation in window environments. The upper left window entitled "SPECIFY SCOPE FOR FORMULATION" is used for model traversal (browsing), where horizontal movement realizes horizontal navigation and a pull-down menu enables vertical immersion for each sub-title (i.e., Application, Subject, and Entity/Relationship). Pertinent objects (application, subject, and entity/relationship) are displayed in the entry fields.

The center window, "FORMULATE QUERY" is for selecting data objects (two types: data item and entity/relationship) into the global query. This way, a user can explicitly include an entity/relationship in a global query for semantic proposes even though no data item will be retrieved from it. The progress of the formulation is displayed in the window on the lower half of the screen. Selection conditions for the query can be specified via the "QUERY FORMULATION IN PROCESS" window.

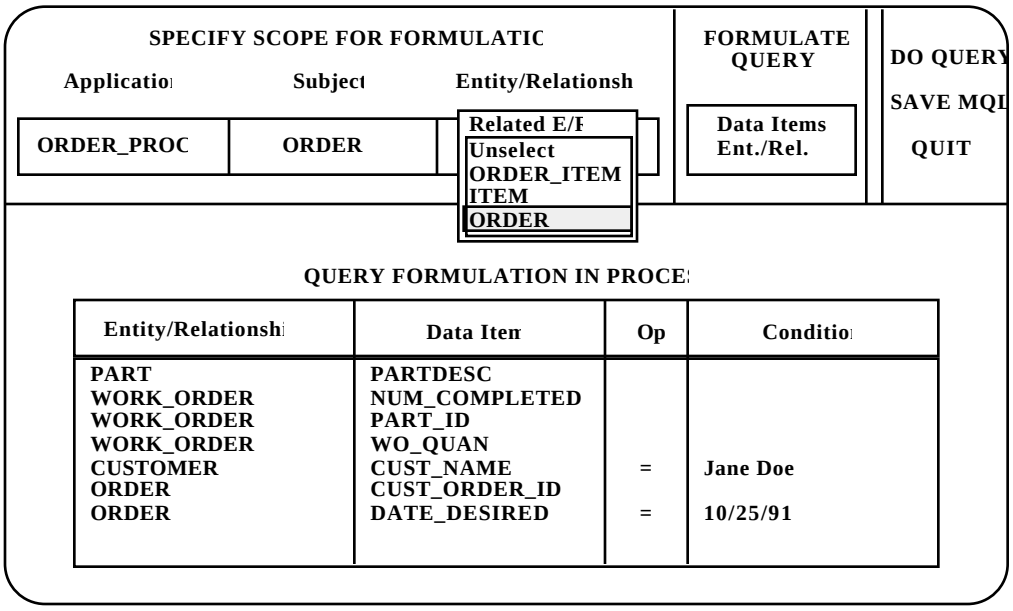


Figure 8 MGQS user interface for global query formulation

The "Do Query" button in the upper right side of the screen will execute the formulated global query. The "Save MQL" button will translate the formulated global query into the syntax of metadatabase query language (see section 4.2) and save it into a file. Lastly, the "QUIT" button will exit from the MGQS system.

The formulated global queries will be decomposed into optimized sub-queries through the global query processor. The query translator then generates local DML code for each sub-query before it is sent to the local system for processing. Finally, as local results arrive, they are interpreted and assembled by the result integrator.

5. AN EMPIRICAL CASE STUDY

5.1 The Heterogeneous Environment

The Computer Integrated Manufacturing (CIM) facility at Rensselaer is used as a test bed for this research. Four functional systems are

employed; namely, a process planning system developed as a dBASE III+ application on an IBM PC/AT, a shop floor control system running under the PC/Oracle DBMS, an order entry system implemented as a Rdb/VAX application, and a product database designed in EXPRESS language and implemented on an object-oriented ROSE platform running on an IBM RS6000. The metadatabase is the fifth system residing on an IBM RS6000. This setup provides a heterogeneous, distributed environment for MGQS prototyping and verification.

The Enterprise Model

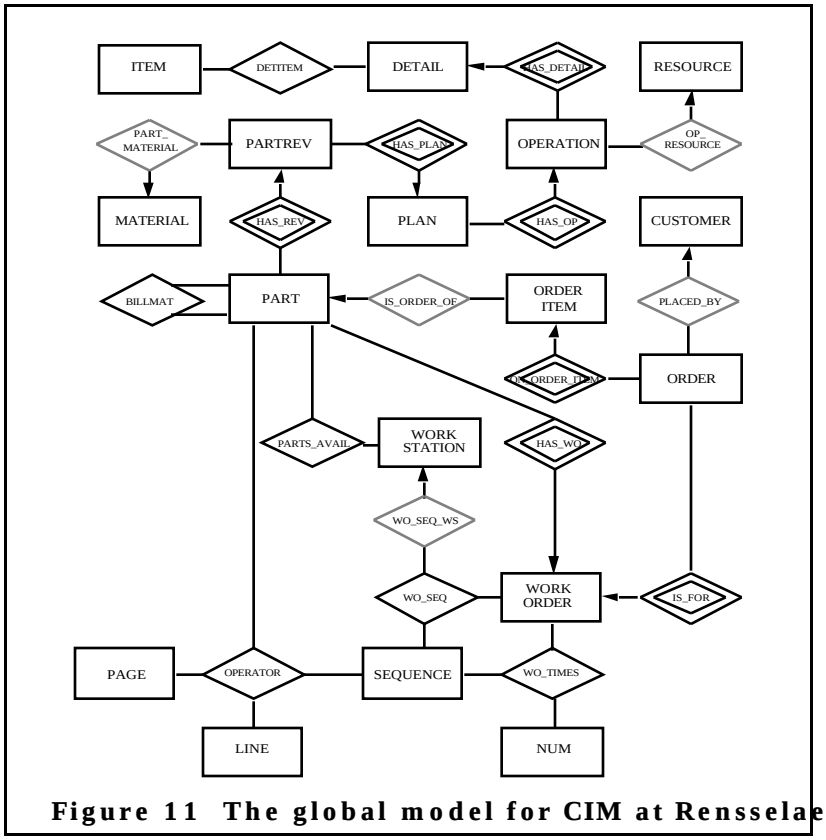


Figure 11 The global model for CIM at Rensselaer

A rigorous modeling and reverse engineering (for paradigm translation) process was engaged to create the enterprise information models and populate the metadatabase. The process, while beyond the scope of this paper, is reported in [29, 30]. This processed yielded a global

data model and other information models created expressly for local systems, and resulted in the enterprise metadata stored into the metadatabase. Thus, all local systems were not disturbed at all by this process; only a separated metadatabase was added to the environment. The global data model for this case is shown in Figure 11.

It is worth noting that not only MGQS does away with imposing data standardization or naming convention on any of the local systems, it also preserves high local autonomy by dealing changes among the local systems dynamically via updating data equivalence, the conversion rules, and other metadata (as pertinent). These changes to the metadatabase require only ordinary database transactions since the GIRD model features metadata independence [28].

5.2 A Global Query Scenario

Consider the following global query example:

"Find the customer order ID, part ID, part description, and quantity completed for Jane Doe's order which has a desired date of 10/25/90."

This request involves data from three application systems; i.e., order entry, shop floor control and process planning. (Note: The user posing this query may not know that the query traverses multiple systems.) The user engages the system through the model-assisted dialog menus and windows to formulate the query while paging through the models. The model traversal starts with selecting an application as an entry point which is very general. As shown in Figure 12, the shaded lines show the particular path followed by the user to formulate this query. Note the relationship between the data items (listed at the bottom of the model) and the

systems in which they are stored (shown at the top). The user does not need to know the information model; it is presented here as an illustration of the global data model network stored in the metadatabase. This network is used by MGQS itself to guide the query formulation process. The items tagged with asterisks have been requested in the query. Also observe that the rounded boxes in the figure represent SUBJECTs and the Squared boxes the corresponding Entities and Relationships of the enterprise model. During the process the user marks the data fields needed along with any conditional statements. (In this query example, DATE_DESIRED = "10/25/90" and CUST_NAME = "Jane Doe") The formulation is completed when all the items are selected and selection conditions are specified.

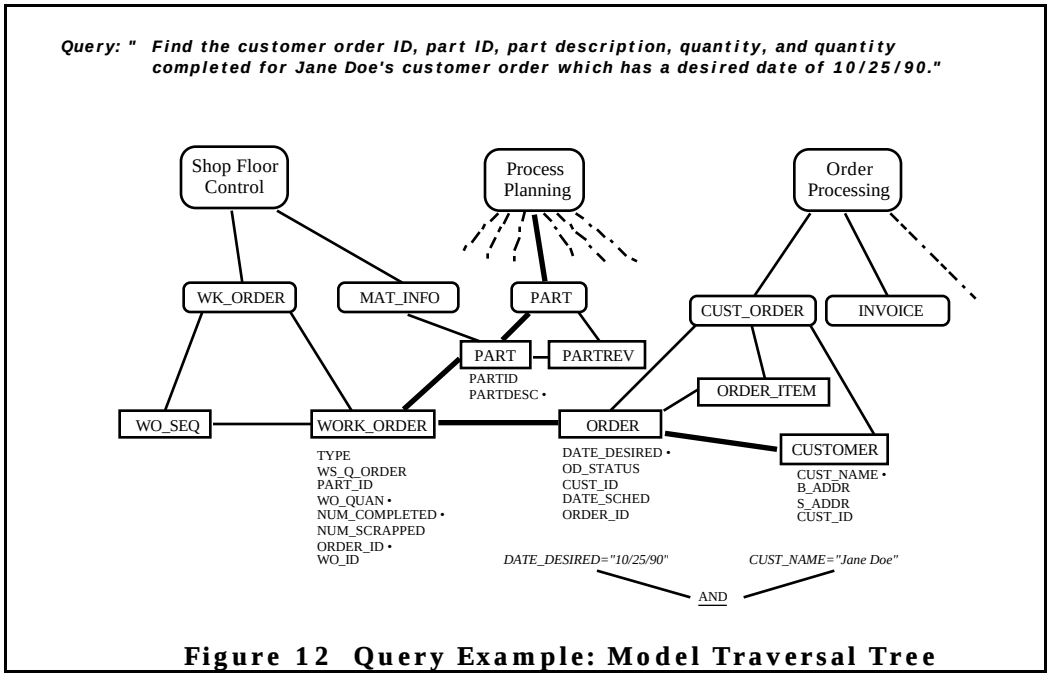


Figure 12 Query Example: Model Traversal Tree

Once the formulation is completed, MGQS will first determine the solution path for the global query from the enterprise model. Equi-join conditions are then inserted to the formulated global query according to the path. The join conditions for this example are:

PART.PARTID = WORK_ORDER.PART_ID

WORK_ORDER.OREDER_ID = ORDER.CUST_ORDER_ID

ORDER.CUST_ID = CUSTOMER.CUST_ID

Note the synonyms; PARTID vs. PART_ID and ORDER_ID vs. CUST_ORDER_ID are identified by the prototype system. Also, PARTID in PART, PART_ID in WORK_ORDER, CUST_ORDER_ID and CUST_ID in ORDER, and CUST_ID in CUSTOMER are not selected during the formulation, therefore MGQS will have to add these items into the global query. Then, MGQS decomposes the global query into a set of optimized, local-bound sub-queries using the implementation models in the metadatabase. Each sub-query requires data retrieval from one and only one local system. MGQS will disseminate these requests using the native query languages of the local systems involved. These local queries are sent out across the network to the respective local systems to be serviced. A rule-based shell system for each and every local systems was included in the prototype to provide data update and other active databases functionalities [4, 29]. This aspect, however is beyond MGQS and is not needed in this discussion here. It suffices the purpose to simply consider that there is a shell at each node serving as the interface between the metadatabase and these local systems.

Each local application system's shell receives a request to process the local query. Upon reception, the sub-query is executed by calling the local DBMS with the file containing the generated local query. The results are then sent back to the MGQS where they were assembled logically and presented to the user (Figure 13).

As shown in Figure 13, the values of the ORDER_ID and CUST_ORDER_ID are coded differently. Data conversion is required in order

to complete the join operation correctly. MGQS calls upon the rule processor to fire the proper conversion rule; contextual knowledge which manipulates raw data or triggers the desired procedures for the conversion before the join.

PART_I	ORDER_I	WO_QUAN	NUM_COM	CUST_ID	CUST_ORDER	DATE_DESIR	CUST_NAME	PART_DESC
PZ1	12345	0	1	JD001	12345_15	10/25/90	Jane Doe	PUZZLE#1
PZ51	12346	0	3	JD001	12346_16	10/25/90	Jane Doe	PUZZLE#51

Figure 13 Final result of the global query

```
FROM OE/PR PART GET PARTID PARTDESC
FROM OE/PR WORK_ORDER GET ORDER_ID WO_QUAN
NUM_COMPLETED
FROM OE/PR ORDER GET DATE_DESIRED
FROM OE/PR CUSTOMER GET CUST_NAME
FOR DATE_DESIRED = "10/25/90" AND
    CUST_NAME = "Jane Doe";
```

Figure 14 MQL formulation

A Test for the Metadatabase Query Language

The same global query example is used for testing the metadatabase query language (MQL). Two formulations for the example using MQL are shown in Figures 14 and 15. Both formulations provide the same result as in Figure 13. Note the formulation in Figure 15 is less precise. The user does not pinpoint the entities and relationships that are involved in the global query. Instead, only the involved applications are specified. The MGQS fills in the necessary information by consulting the contents of the metadatabase. Minimally, a user will only need to specify the data item(s) and the necessary selection condition(s) for a global query. The rest of the query processing is the same as described before.

```
   FROM APPLICATION  PROCESS_PLAN GET PARTID PARTDESC
   FROM APPLICATION  SHOP_FLOOR GET WO_QUAN
NUM_COMPLETED
   FROM APPLICATION  ORDER_ENTRY GET DATE_DESIRED
CUST_NAME
   FOR DATE_DESIRED = "10/25/90" AND
      CUST_NAME = "Jane Doe";
```

Figure 15 MQL formulation

5.3 Discussion

It is worth noting that the OES was originally implemented as a VAX/VMS file system. New OES using the Rdb, a relational DBMS on Micro VAX, was designed and built after the MGQS prototype had been functionalized. The adaptation to the new system for MGQS was done rather quickly. First, we removed the old OES models from the contents of the metadatabase. Then the TSER models of the OES, obtained from the top-down modeling effort, were incorporated automatically into the metadatabase (i.e., consolidating the new model with the existing enterprise model and populating the metadatabase with the new models). Finally a customized code generator for Rdo, a DML for Rdb, is programmed. The total effort for the changes essentially was completed when the code-generator was completed. This situation illustrates the idea of changes under metadata independence as mentioned before.

In the whole, the CIM facility at Rensselaer provides an environment with a reasonable complexity to test the prototype MGQS and the model-assistance concept. All major objectives (i.e., information sharing, local system autonomy, and model assistance) of the MGQS are achieved and proven to be feasible with the prototype system. Most of the envisioned

functionalities (with the exception of ambiguity checking and derived item querying) are implemented. In addition, the data conversion capability has demonstrated a use of the contextual knowledge and rule processor for providing on-line intelligence. This same method can be applied to the function of ambiguity checking and derived data querying.

6 ANALYSIS

The MGQS approach is justified in this section through a comparison with previous results with respect to the objectives (Section 2.1) of information sharing in heterogeneous distributed environments. A comment on the generalizability of the MGQS results is also provided.

6.1 Metadata Modeling vs. Schema Integration

A key element in conventional distributed databases is schema integration. All databases are logically structured and controlled under a single integrated schema in a homogeneous environment [11, 43, 53]. The user would be able to share information across all databases as if there were only one classical centralized database. Therefore, local transparency and conflict resolution are achieved through enforcing an integrated schema under a single data model. The primary problem with this approach, however, is lack of local system autonomy and adaptability.

To illustrate, consider the following example shown in Figure 16. In this example SNO and SID are two synonyms of the same logical attribute (one in data object SUPPLY of database PJ at site A and the other in data object SUPPLIER of database S at side B). Further, PNO in PART and the PNO in SUPPLY are structured differently with different domains CODE1 and CODE2, respectively. The integrated schema could be developed as shown in Figure 17 (assuming a relational system). In the integrated

schema, the two conflicting definitions are reconciled by changing attribute name SID to SNO in SUPPLIER and the value domains of PNO in both SUPPLY and PART to type integer. The local systems will recompile and reload so as to implement the changes according to the integrated schema.

Figure 18 shows the global model that MGQS would employ for this example. Note that each data item in the global model is assigned with a unique item code For instance, the two PNO's in PART and SUPPLY are logically the same attributes with the same name but

Site	Database	Data Objects	Attributes	Format/Domain
A	PJ	PROJECT	JNO	INTEGER
			J_LOC	CHAR(40)
		SUPPLY	SNO	INTEGER
			PNO	CODE2
			JNO	INTEGER
B	S	SUPPLIER	SID	INTEGER
			S_LOC	CHAR(40)
C	P	PART	PNO	CODE1
			PNAME	CHAR(20)
			LENGHT	REAL

Figure 16 An example of three systems A, B, and C

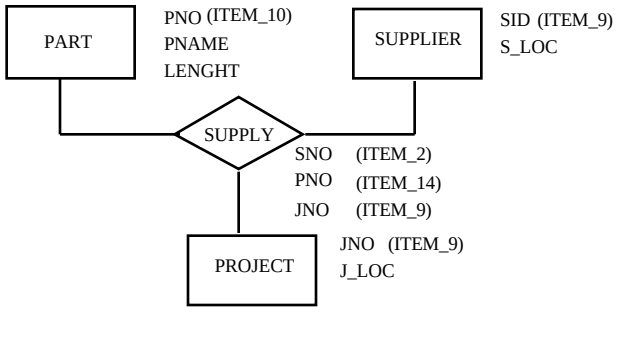
```
CREATE TABLE PROJECT (
    JNO  INTEGER,
    J_LOC  CHAR(40));
CREATE TABLE SUPPLY (
    SNO  INTEGER,
    PNO  INTEGER,
    JNO  INTEGER);
```

```
CREATE TABLE PART(
    PNO  INTEGER,
    PNAME  CHAR(20),
    LENGTH  REAL);
CREATE TABLE SUPPLIER (
    SNO  INTEGER,
    S_LOC  CHAR(4O));
```

Figure 17 Integrated schema for the example using DDBMS approach

reside in different local systems, and hence they have different item codes. However, the attribute JNO in the database PJ is only a single data item since, albeit shared by both SUPPLY and PROJECT, it is the same attribute with the same name in the same system, therefore they have the same

item code. The fact that PNO in PART and SUPPLY are coded differently is represented as a tuple in Equivalent along with the conversion rules Rule1 and Rule2. Synonyms SID and SNO are also reflected in Equivalent. Therefore, no change is required of the local systems using this method while any names could be used by users in any systems to address the attributes and obtain globally consolidated results. The same process is utilized to effect equivalent with respect to types, semantic formats (e.g., dates) and user-dependent presentation. Note that the process does not require an integrated global schema. Concerning adaptability such as modifying, adding or deleting (local) models, any such changes to the enterprise models can be simply handled as metadata transactions against the metadatabase. without affecting its schema (the GIRD model). This, as mentioned in section 3.1, is the significance of metadata independence. Most of previous results, some of which avail the similar local system autonomy as discussed in the above example, do not satisfy metadata independence.



Equivalen

Itemcode	Eqitemcode	reverse_by	convert_by
ITEM_10	ITEM_14	Rule1	Rule2
ITEM_2	ITEM_9	NULL	NULL

RULE

Rname	rtype	description
Rule1	M	convert item_14 to item_10
Rule2	M	convert item_10 to item_14

Figure 18 Global model for the example using the MGQS

6.2 MQL vs. MSQL

The MSQL [35, 36] facility is an extension of SQL for manipulating data managed by the multidatabase. The major results include (1) additional commands for multiple databases and (2) enhanced naming convention for data objects. Applying MSQL to the same global query example in Figure 12 would yield the following result:

```
Create Multidatabase  CIM (SHOPFLOOR PROCESS_PLAN ORDER)
Use     CIM
Select SHOPFLOOR.WORK_ORDER.PART_ID,
       SHOPFLOOR.WORK_ORDER.ORDER_ID, SHOPFLOOR.WORK_ORDER.QUAN,
       SHOPFLOOR.WORK_ORDER.COMPLETED, PROCESS_PLAN.PART.PARTDES,
       ORDER.ORDER.DATE_DESIRED, ORDER.CUSTOMER.CUST_NAME
From   SHOPFLOOR.WORK_ORDER, PROCESS_PLAN..ORDER, PROCESS_PLAN.PART,
       ORDER.CUSTOMER
Where      PROCESS_PLAN.PART.PARTID = SHOPFLOOR.WORK_ORDER.PART_ID
And    SHOPFLOOR.WORK_ORDER.ORDER_ID = ORDER.ORDER.ORDER_ID
And    ORDER.CUSTOMER..CUST_ID = ORDER.ORDER.CUST_ID
```

And ORDER.ORDER.DATE_DESIRED = “10/25/93”;

The above example is evidently more complicated than using MQL (see Figure 14, 15). The reason is simple: MSQL does not utilize a metadata facility for global query formulation. Users have to provide all the technical details including precise location of data items, data equivalence, and join conditions. Furthermore, two equivalent data items, PARTID in PART and PART_ID in WORK_ORDER, in a join condition must have the same format in order for MSQL to integrate the results.

6.3 MGQS vs. Multibase

Multibase [15, 50] is one of the pioneer research efforts and arguably the most well known practical system in the field of heterogeneous, distributed DBMS. A single query language (DAPLEX) based on the functional data model is employed for information retrieval. Preserving local system autonomy and resolving data incompatibilities are the major design objectives for the Multibase. Thus, it is a natural reference point for MGQS to be compared with.

Multibase has three levels of schema: (1) a global schema at the top level, (2) an integration schema and one local schema for each local system at the middle level, and (3) one local host schema per local database at the bottom level. The global schema, local schemata, and the integration schema are all defined according to and in terms of the functional data model. Each of the local host schemata is translated into a local schema and the integration schema is used to describe the so-called integration database containing information needed for integrating results from local databases (i.e., information about mapping between conflicting data

definitions). The local schemata and integration schema are then mapped into global schema.

To provide a single query language for information retrieval from heterogeneous systems, Multibase has to synchronize multiple data definitions at the global schema level. Also, multiple data definitions for the logically identical data are not allowed. This is the classical case of integrated schema as discussed in Section 6.1. In contrast, MGQS supports a single global query environment by consulting the metadatabase instead of imposing restrictions via the global model. Multiple data definitions are allowed and they are stored in the metadatabase along with the conversion methods.

6.4 MGQS vs. Other HD_DBMSs

Other heterogeneous, distributed DBMSs such as Federated databases [46] have proposed similar three level schema structures as Multibase to achieve local system autonomy. The basic difference from the latter is, in order to promote an open architecture, these approaches do not enforce a single global model. Instead, several logically related sub-systems are grouped together with an integrated model (referred to as external schema) developed for them. Multiple external schemata are coexisting in the global system. This approach provides more flexibility for integration design and also reduces the scope of schema integration by cutting down the number of sub-systems involved in an integrated model. However, since each sub-system may participate in multiple external schemata which are custom developed, evolving the overall structure, such as incorporating a new sub-system to the environment is still fundamentally more complicated.

Moreover, there is no metadata management facilities used. Global[2] query formulation relies, therefore, on an extended query language and the user's technical knowledge of the sub-systems and their external schemata. A minimal extension of the language is to allow relation names to be qualified with database names in case of naming conflicts. To address an attribute, the user has to know precisely which relation of what database the attribute belongs to. For instance, PNO of PART is referred to as P.PART.PNO. Thus local system transparency is not completely supported in these approaches.

6.5 MGQS vs. Other Metadata Supported Systems

Increasingly, more global query systems have come to the same conclusion of utilizing metadata to resolve some well know problems in query processing. The driving issues include semantics, optimization path selection, and result integration; which correspond respectively to steps 1.3, 2 (particularly 2.3), and 6 of the MGQS definitional algorithm in Section 2.2.1. Some metadata management facilities may also be used in the forms of a knowledge base or a repository. Their main difference from the MGQS approach lies in the scope of metadata and the metadata independency of the management facilities. A prime example is the Composite Information Systems (CIS) [49, 55, 49] of MIT. It is one of the first works that emphasize the significance of metadata and use it to resolve some semantic conflicts in query translation and results integration. Since there is no formal knowledge methods included in its metadata facility used, the scope and functionality would be affected. In addition, CIS assumes all local systems and the global schema being relational.

The Carnot approach of MCC (Microelectronics and Computer Technology Corporation) [14] differs from the CIS approach in its global schema construction and the use of metadata representation model; namely the Cyc Knowledgebase. This approach makes the global schema much easier to construct and maintain. However, as stated in [14], there are two major problems that the Carnot approach has not yet resolved; i.e., (1) how to integrate the results returned from sub-system queries, and (2) how to develop a graphical entity-relationship representation of the global schema and an intelligent interface for specifying queries. Both of whicgh are direct results of the Cyc knowledgebase design. The MGQS approach facilitates both, due in part to the scope and the structure of the metadatabase (see Figures 2 and 3, and Sections 3.5 and 6.1).

In contrast to CIS and Carnot, the recently reported KIM (Knowledge-base Information Manager) Query System [16] provides an iconic ER based user interface for global query formulation, but does not seem to address such issues as system evolution, optimization path and results integration. Again, although KIM features a repository similar to Carnot, its design does not support a full vision of enterprise metadata.

6.6 Conceptual Evaluation of the MGQS Approach

The model-assisted global query approach provides several important functionalities for sharing information in heterogeneous distributed environments. Based on the discussions in the above sections, these properties are categorized into five areas, as discussed below.

Maintain local system autonomy using the metadatabase

As mentioned above, MGQS employs on the metadatabase for identifying and resolving the heterogeneities among these systems.

Instead of imposing an integrated schema over all the local systems, the metadatabase approach allows the local systems to keep their environments insulated for their own as well as to control their own data definitions and everything else. The metadatabase and local systems are completely independent bubbles whose operations do not rely, nor infringe, on each other; unlike the integrated schema environments. The functional model, structural model, implementation model and the associations among these models of each local system are independently represented, consolidated, and stored in the metadatabase. Thus, their manipulation and management are as easy as any ordinary data in a database.

Allow direct and visual query formulation from enterprise metadata

A global query is formulated by selecting the data constructs while browsing the global enterprise information models. Regardless of the entry point chosen, the user will eventually lead to the needed data constructs by paging through the global model without noticing the boundary of local systems. This is a high level of direct manipulation (of information objects) that researchers have advocated for cognitive user interface. The constructs used (see Figure 2) are clearly compatible with graphical representation and hence supports immediately visual methods for query formulation.

Provide on-line intelligence and assistance for challenging tasks such as semantics, path selection, and result integration

MGQS provides on-line knowledge to facilitate both global query formulation and processing. During query formulation, the pertinent

information contents and semantics of the heterogeneous systems are either provided interactively to the user, or MGQS utilizes them to automate certain tasks for the user (see sections 2 and 3). The on-line knowledge provided to the global query processing is transparent to the user. They include the automatic decomposition and optimization of global query and the recognition and conversion of equivalent data resources across local systems. Combined, the approach provides a broad range of metadata support covering the entirely of the definition discussed in Section 2.2. They are sufficient for certain difficult tasks, including semantics, path selection for query optimization, and result integration using data equivalence knowledge (see above comparisons). Since the metadatabase contains both enterprise knowledge resources and data models, its extent of metadata is unique, and capable of providing significant on-line intelligence and assistance.

Include rule-based knowledge processing for extensibility

MGQS acquires contextual knowledge from the knowledge model of the metadatabase to provide on-line intelligence for the global query operation. Two representative examples of contextual knowledge are business rules which describe the intended use of data and operating rules which facilitate decision processes across systems. Global queries involving derived items is another example on the use of contextual knowledge. A rule base model and an inference engine are part of the metadatabase system. New rules concerning MGQS can be relatively easily incorporated into its architecture and execution model.

Achieve open system architecture through metadata independence

The MGQS approach provides an open system architecture which is flexible for adding, deleting, or modifying a local system in the integrated enterprise. The property is referred as metadata independence in Section 3.1. For instance, to remove the shop floor control system from the CIM enterprise in section 5, no change is required of the MGQS. The metadata pertaining to the shop floor system will be removed from the metadatabase through ordinary transactions (i.e., deleting tuples from the meta-relations). Accepting a new sub-system to the integrated enterprise requires primarily modeling effort, which is no more than those required by conventional approaches. However, the addition of this new model to the existing global model is merely a matter of performing, again, ordinary metadatabase addition transactions. A documentation of this modeling process is provided, e.g., in [29, 31].

6.7 The Generalizability of the MGQS Approach

How much of the MGQS approach can be utilized by other systems than the prototype at Rensselaer which makes heavy use of a particular metadatabase and a particular modeling method TSER? The answer turns out to be correlated in steps with the development from Section 2 to Section 4. The conceptual model in Section 2 is most general and generic. Any system may be able to develop its own metadata facility (be it a knowledge base, a repository, or a metadatabase) utilizing the metadata requirements specified in the model. At the next level, the particular MGQS execution model is based on the particular metadatabase. Although its design is rooted in TSER, the GIRD model is simply a specification for

creation of the metadatabase. Therefore, the model maybe implemented in any system supporting the usual relational-compatible facilities. The third level, the most specific, involves TSER modeling and the metadatabase management system. The full repetition of the results reported in Sections 5 and 6 will require all elements in Sections 2, 3, and 4. However, these elements are available in the open literature and can be adopted by the general readership.

7 CONCLUSIONS AND CONTRIBUTIONS

The Model-Assisted Global Query System accomplishes better information sharing in heterogeneous, distributed environments than previous query systems. The goals of the approach and its conceptual model were discussed in sections 1 and 2, and justified in sections 5 and 6; while the execution methods were established in sections 3 and 4.

In particular, MGQS contributes a direct method to end user query formulation through on-line assistance using metadata (section 2). It allows the user to articulate directly in terms of information models with which they are familiar. The pertinent information contents and data semantics of the heterogeneous multiple systems are provided interactively to the user, thereby further alleviating the technical complexities and semantic ambiguity during formulation. In a similar manner, some technical support is also afforded, including diagnosis and feedback of query formulation according to the business rules and other contextual knowledge in the system. This direct method contributes in its own right to the conceptual foundations of graphical and interactive user interface technology.

New methods that utilize the on-line knowledge (or metadata) for major global query processing tasks are also developed. They encompass the areas of global query optimization and decomposition, query translation, and result integration. The knowledge needed for these tasks (e.g., implied data items, shortest solution path, and join conditions; local system access paths; query construction across local database schemata; and data equivalency and conversion) is automatically derived from the metadatabase. Without such on-line intelligence and assistance, the required knowledge would have to either be supplied by users at run/compile-time, or be predetermined at design time through schema integration and other standardization approaches.

This work has resolved some interoperability issues of heterogeneous multiple information systems through offering an alternative to the conventional approaches which rely on schema integration. Schema integration is a major source of technical complexity of heterogeneous distributed DBMS at both design-time (efforts and restrictions) and run-time (mappings and architectural overheads). Additional knowledge such as conflicting or alternating data definitions and their resolutions is also supported and stored in the metadatabase for MGQS. Conversion is done in real time through a rule processor firing conversion rules. As a result, the only global modeling effort required of this approach is the development of the enterprise information model itself. It is still challenging, especially concerning knowledge acquisition; but it nevertheless avoids the excessive complexity of integration at the schemata level that would ensue on the enterprise model in the case of conventional approaches.

This research has also identified and characterized for the first time a definitive model for the notion of "on-line intelligence and assistance" in end user query interface and query processing in term of the enterprise metadata. It is also shown through MQL, its query language, that programming query languages can also benefit from the on-line intelligence and assistance. As we have shown in Section 4, the syntax and the technical details of MQL are significantly improved, compared to other existing global query languages. The concept and new methods were tested with the prototype system using the CIM facility at Rensselaer.

Further research are currently underway to enhance MQL and MGQS methods, especially by incorporating additional rule-oriented capabilities into the system. MGQS is also being employed in a seamless way to facilitate global data management and event-based information flows management in the metadatabase research. New progress might come from applying the basic methods to other information modeling paradigms (than TSER). Finally, information visualization [32] will be a natural extension to MGQS, where an agent may be developed to personalize the metadatabase and a visual/virtual environment can replace the GUI-based user interface used presently.

REFERENCES

1. Afsarmanesh, Hamideh, and D. McLeod, "The 3DIS: An Extensible Object-Oriented Information Management Environment," ACM Transactions on Information Systems, October 1989, Vol. 7, No. 4, pp. 339-377.

2. Angelaccio, M., T. Catarci, and G. Santucci, "QBD*: A Graphical Query Language with Recursion," IEEE Transactions on Software Engineering, Vol. 16, No. 10, pp. 1150-1163, 1990.

3. Azmoodeh, M., "BRMQ: A Database Interface Facility based on Graph Traversals and Extended Relationships on Groups of Entities," The Computer Journal, Vol. 33, No. 1, pp. 31-39, 1990.

4. Babin, G. "Adaptiveness in Information Systems Integration," unpublished Ph.D. Dissertation, Decision Sciences and Engineering Systems Department, Rensselaer Polytechnic Institute, Troy, N.Y., 1993.

5. Benjamin, A.J., and K.M. Lew, "A Visual Tool for Managing Relational Databases," Proceedings IEEE International Conference on Data Engineering, pp. 661-668, 1986.

6. Bernstein, P. et al., "Query Processing in a System for Distributed Databases (SDD-1)," ACM Transactions on Database Systems, 1981.

7. Batini, C., M. Lenzerini, and S.B. Navathe, "A Comparative Analysis of Methodologies for Database Schema Integration," ACM Computing Surveys, Vol. 18, No. 4, 1986.

8. Blaser, A., Natural Language at the Computer: Scientific Symosium on Syntax and Semantics for Text Processing and Man-Machine-Communication, New York: Springer-Verlog, A. Blaser, (ed.,) 1988.

9. Bouziane, M. "Metadata Modeling and Management," unpublished Ph.D. Dissertation, Computer Science Department, Rensselaer Polytechnic Institute, Troy, N.Y., 1991.

10. Bunt, H.C., " Natural Language Communication with Computers: Some Problems, Perspectives, and New Directions," Human-Computer Interaction: Psychonomic Aspects, G.C. Veer and G. Mulder (ed.,) Springer-Verlag, pp. 406-442, 1988.

11. Cardenas A.F., "Heterogeneous Distributed Database Management: The HD-DBMS," Proceedings of the IEEE, Vol. 75, No. 5, 1987.

12. Cheung, W. "The Model-assisted Global Query System," unpublished Ph.D. Dissertation, Computer Science Department, Rensselaer Polytechnic Institute, Troy, N.Y., 1991.

13. Chung, C.W., "DATAPLEX: An Access to Heterogeneous Distributed Databases," Communications of the ACM, Vol. 33, No. 1, pp.70-80, January 1990.

14. Collet, C., M.N. Huhns, and W.M. Shen, "Ressource Integration Using a Large Knowledge Base in Carnot," IEEE Computer, pp. 55-62, December 1991.

15. Dayal, U., and H. Hwang, "View Definition and Generalization for Database Integration in MULTIBASE: A System for Heterogeneous Distributed Databases," IEEE Transactions on Software Engineering, Vol. SE-10, No. 6, pp. 628-644, 1984.

16. Ferrara, F.M., " The KIM Query System: An Iconic Interface for the Unigied Access to Distributed Multimedia Databases," ACM SIGCHI, Vol. 26, No. 3, pp. 30-39, July 1994

17. Gardiner, M.M., and B. Christie, Applying Cognitive Psychology to User-Interface Design, John Wiley & Sons, 1987.

18. Gould, J.D. and C. Lewis, "Designing for Usability: Key Principles and What Designer Think," Communications of the ACM, Vol.28, No. 3, pp. 300-311, 1985.

19. Greenberg, S. and I.H. Witten, "Adaptive Personalized Interfaces - A Question of Viability," Behavior and Information Technology, Vol. 4, No. 1, pp. 31-45, 1985.

20. Gyssens, M. J. Paredaens, and D.V. Gucht, "A Graph-Oriented Object Model for Database End-User Interfaces," ACM, pp. 24-33, 1990.

21. Hartson, H.R. and D. Hix, "Human-Computer Interface Development: concepts and Systems for Its Management," ACM Computing Surveys, Vol. 21, N0. 1, pp.5-92, March 1989.

22. Herot, C., "Spatial Management of Data," ACM Transactions on Database Systems, Vol. 5, No. 4, pp. 493-513, 1980.

23. Hirschman, L. "Natural Language Interfaces for Large-Scale Information Processing," Integration of Information Systems: Bridging Heterogeneous Databases, A. Gupta, (ed.,) IEEE Press, New York, pp.308-314, 1989.

24. Hsu, C. "Structured Databases Systems Analysis and Design Through Entity-Relationship Approach," Proceedings 4th International Conference on Entity Relationship Approach, pp. 56-63, 1985.

25. Hsu, C., A. Perry, M. Bouziane, and W. Cheung, "TSER: A Data Modeling System Using the Two-Stage Entity-Relationship Approach," Proceedings 6th International Conference on Entity Relationship Approach, New York, pp. 461-478, November 1987a.

26. Hsu, C., and C. Skevington, "Integration of Data and Knowledge Model in Manufacturing Enterprises: A Conceptual Framework," Manufacturing Systems, Vol. 6:4, pp.277-285, 1987b.

27. Hsu, C., and L. Rattner, " Information Modeling for Computerized Manufacturing," IEEE Transactions on Systems, Man, and Cybernetics, Vol. 20, No.4, pp. 758-776, 1990b.

28. Hsu, C., M. Bouziane, L. Rattner, and L. Yee, "Information Resources Management in Heterogeneous, Distributed Environment: A Metadatabase Approach," IEEE Transactions on Software Engineering, Vol. SE-17, No. 6, June pp.604-625, 1991.

29. Hsu, C., G. Babin, M. Bouziane, W. Cheung, L. Rattner, and L. Yee, "Metadatabase Modeling for Enterprise Information Integration," Journal of Systems Integration, Vol.2 No. 1, pp.5-37, 1992.

30. Hsu, C., Y. Tao, M. Bouziane, G. Babin, "Paradigm Translations in Integrating manufacturing Inforamtion Using a Meta-model: The TSER Appraoch," Journal of Information Systems Engineering, France, Vol. 1, No. 3, pp325-352, 1993a..

31. Hsu, C., G.Babin, M. Bouziane, W. Cheung, L. Rattner, A. Rubenstein, and L. Yee, "The Metadatabase Approach to Integrating and managing Manufacturing Inforamtion Systems," Journal of Intelligent Manufacturing, forthcoming.

32. Hsu, C. and L. Yee, "Model-Based Visualization for Enterprise Infromation Management," Proceedings of the 4th Annual Conference on AI, Simulation and Planning in High Autonomous Systems, Tucson, Arizona, pp.324-327, September 1993.

33. Hutchins, E.L., J.D. Jollan, and D.A. Norman, "Direct Manipulation Interfaces," User Centered System Design, D.A. Norman and S.W. Draper (ed.,) Hillsdale NJ: Lawrence Erlbaum Associates, pp. 118-123, 1986.

34. Krishnamurthy, R., W. Litwin, and W. Kent, "Language Features for Interoperability of Databases with Schematic Discrepancies," Proceedings 1991 ACM SIGMOD International Conference on Management of Data, pp.40-49, 1991.

35. Litwin, W., A. Abdellatif, a. Zeroual, and B. Nicolas, "MSQL: A Mulitdatabase Language," Information Sciences, Vol. 49, pp.59-101, 1989.

36. Litwin, W., L. Mark, and N. Roussopoulos, "Interoperability of Multiple Autonomous Databases," ACM Computing Surveys, Vol. 22, No. 3, pp. 267-293, 1990.

37. Lodding, K.N., "Iconic Interfacing," IEEE Computer Graphics and Applications, Vol. 3, No. 2, pp. 11-20, 1983.

38. Motro, A., "FLEX: A Tolerant and Cooperative User Interface to Databases," IEEE Transactions on Knowledge and Data Engineering, Vol. 2, No. 2, pp.231-246, June 1990.

39. Nilan, Michael. S., "Cognitive Space - Using Virtual Reality For Large Information Resource Management Problems," Journal of Communication, Autumn 1992, Vol. 42, pp. 115-135.

40. Norcio A.F., and J. Stanley, "Adaptive Human-Computer Interfaces: A Literature Survey and Perspective," IEEE Transactions on Systems, Man, and Cybernetics, Vol. 19, No. 2, pp.399-408, March/April 1989.

41. Rattner, L., "Infromation Requirements for Integrated Manufacturing Planning and Control: A Theoretical Model," unpublished Ph.D. Dissertation, Department of Decision Sciences and Engineering Systems, Rensselaer Polytechnic Institute, Troy, NY, 1990.

42. Reddy, M.P., B.E. Prasad, P.G. Reddy, "Query Processing in Heterogeneous Distributed Database Management Systems," Integration of Information Systems: Bridging Heterogeneous Databases, A. Gupta (ed.,) IEEE Press, pp.264-277, 1989.

43. Rem, S., "Guest Editor's Introduction: Heterogeneous Distributed Database Systems," IEEE Computer, Vol. 24, No. 12, December 1991, pp. 7- 11.

44. Rich, E., "Natural-Language Interfaces," IEEE Computer, pp.39-47, September 1984.

45. Robertson, George C., S. K. Card, and J. D. Mackinlay, "Information Visualization Using 3D Interactive Animation," Communications of the ACM, Vol. 36, No. 4, pp. 56-71, 1993.

46. Sheth A.P., and J.A. Larson, "Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases," ACM Computing Surveys, Vol. 22, No. 3, pp.183-236, 1990.

47. Shipman, D.W., "The Functional Data Language DAPLEX," ACM Transactions on Database Systems, Vol. 6, No. 1, pp.140-173, 1981.

48. Shyy, Y.M., and S.Y.W. Su, "K: A High-level Knowledge Base Programming Language for Adveanced Database Application," Proceedings 1991 ACM SIGMOD International Conference on Management of Data, pp.338-346, 1991.

49. Siegel, M., and S. Madnick, "A Metadata Approach to Resolving Semantic Conflicts," Proceedings of the 17th International Conference on Very Large Data Bases, pp. 133-146, September, 1991.

50. Smith, J.M., P.A. Bernstein, U. Dayal. N. Goodman, T. Landers, K.W.T. Lin, and E. Wong, "Multibase Integrating Heterogeneous Distributed Database Systems," Proceedings of AFIPS NCC, Vol. 50, pp. 487-499, 1981.

51. Stonebraker, M., "Introduction for User Interfaces," Readings of Database Systems, M. Stonebraker (ed.,) Morgan Kaufmann Publishers, Inc. San Mateo, California, pp.337-339, 1988.

52. Stonebraker, M., and J. Kalash, "TIMBER: A Sophisticated Relation Browser," Proceedings of Eighth International Conference on Very Large Data Bases, pp. 1-10, September 1982.

53. Templeton, S. Fox, and B. Hartman, "Heterogeneous Distributed Database Systems for Production Use," ACM Computing Surveys, Vol. 22, No. 3, pp. 237-266, 1990.

54. Wang, R., and S. Madnick, "Facilitating Connectivity in Composite Information Systems," ACM Database, pp. 38-46, Fall 1989.

55. Wang, R., and S. Madnick, "A Ploygen Model for Heterogeneous Database Systems: The Source Tagging Perspective," Proceedings of 16th International Conference on Very Large Data Bases, pp. 519-538, September 1990.

56. Zloof, M.M., "Query-by-Example: A Data Base Language," IBM Systems Journal, Vol. 4, pp.324-343, 1977.