The Metadatabase Project at Rensselaer

Cheng Hsu
Department of Decision Sciences and Engineering Systems
Rensselaer Polytechnic Institute, Troy, NY 12180-3590
(518) 276-6847 Bitnet: USERCHBY@RPITSMTS

Abstract

The Metadatabase project is a multi-year research effort at Rensselaer Polytechnic Institute. Sponsored by industry (ALCOA, DEC, GE, GM, IBM and others) through Rensselaer's Computer Integrated Manufacturing Program, this project seeks to develop novel concepts, methods and techniques for achieving information integration across major functional systems pertaining to computerized manufacturing enterprises. Thus, the metadatabase model encompasses the generic tasks of heterogeneous, distributed and autonomous databases administration, but also includes information resources management and integration of concurrent (functional) systems. The model entails (1) an integrated data and knowledge modeling and representation method; (2) an online kernel (the metadatabase) for information modeling and management; (3) metadatabase assisted global query formulation and processing; (4) a concurrent architectural whereby global synergies are achieved through (distributed) metadata management rather than synchronization of (distributed) database processing; and (5) a theory of information requirements for integration. A metadatabase prototype was recently demonstrated to the industrial sponsors. The basic concept of the metadatabase model is discussed in this paper.

1. The Metadatabase Model.

At an enterprise level, the objective of information integration extends beyond technical interfacing of computing environments and envisages functional synergies across the manufacturing enterprise as a whole, including business and engineering as well as production systems [1, 4, 5]. Therefore, this integration problem presents several unique challenges over and above the generic issues of heterogeneous, distributed and autonomous databases. (See [9, 10, 13] for discussions of some recent works on the generic problem, and [8, 15, 16] for representative efforts on the manufacturing front.)

Foremost is the need to incorporate contextual knowledge with databases. This need stems from two basic facts: (1) computerized manufacturing enterprises often include various knowledge-based systems that are an integral part of the overall integration, and (2) the functional contexts in which individual database systems contribute to enterprise-wide synergy must be sufficiently represented. In general, at least three major classes of contextual knowledge should be considered:

(1) operating knowledge (business rules) concerning individual data models or functional (sub-)systems (e.g., triggers, processes, and semantic constraints); (2) control knowledge for sequential systems interactions, including data transfer rules and global equivalence knowledge for all data items pertaining to the same logical object but implemented differently; and (3) decision knowledge for parallel systems interactions, such as global decision processes and their implementation into information flows, localized decision rules, and control procedures. (See [11] for a full discussion on the last two classes of knowledge.)

Other challenges include acute needs for supporting heterogeneity, concurrency, and end-user computing due to the scope of integration. In particular, the characteristics of heterogeneity range from calling for relational type of environments (for, e.g., business and production planning applications) be coupled with object-oriented type of systems (for engineering or even shop floor control), all the way to demanding rule-based knowledge systems be adjoined with databases irrespective of their models. In a similar way, the requirements of concurrency and end-user computing encompass all major functions and decision processes pertaining to the synergies of engineering, production, and business applications. These issues are further compounded by the large volume of data processing and communications (typically millions of transactions in a modern manufacturing facility), which tend to render global serialization impractical.

Together, these challenges give rise to a complexity beyond the traditional concerns of data and knowledge engineering technology. Novel approaches focusing on simplifying the complexity itself as well as dealing with it are needed. Toward this end, a metadatabase model is developed at Rensselaer through the industry-sponsored Computer-Integrated Manufacturing Program [4, 5]. This model seeks to develop new concepts, methods and techniques for achieving information integration at the enterprise level. It extends the notion of metadata into an online kernel of enterprise information facilitating the integration tasks at all major levels. More formally, a metadatabase in its own right is a (new type of) database integrated with a knowledge base. It aims to satisfy a diverse set of applications across the enterprise with a unified architecture. First, there are enterprise users such as managers and systems developers who either need to maintain or want to reap benefits from enterprise information resources. The metadatabase supports these functions. In addition, the metadatabase is a repository of intelligence into which all user groups in different functional (sub-)systems or databases can tap for decision support or global query formation and processing. Last, but not least, in an active mode whereby all systems interact with each other, the metadatabase implements the above three classes of knowledge and is shared among (parallel) processes and automated systems to facilitate information integration. The metadatabase also has other characteristics of application databases, including metadata integrity and independence. Based on this model, an integration architecture is illustrated in Figure 1.

2. Current Capabilities.

Three basic elements are necessary for the metadatabase environment: (1) a logical model to define and represent the

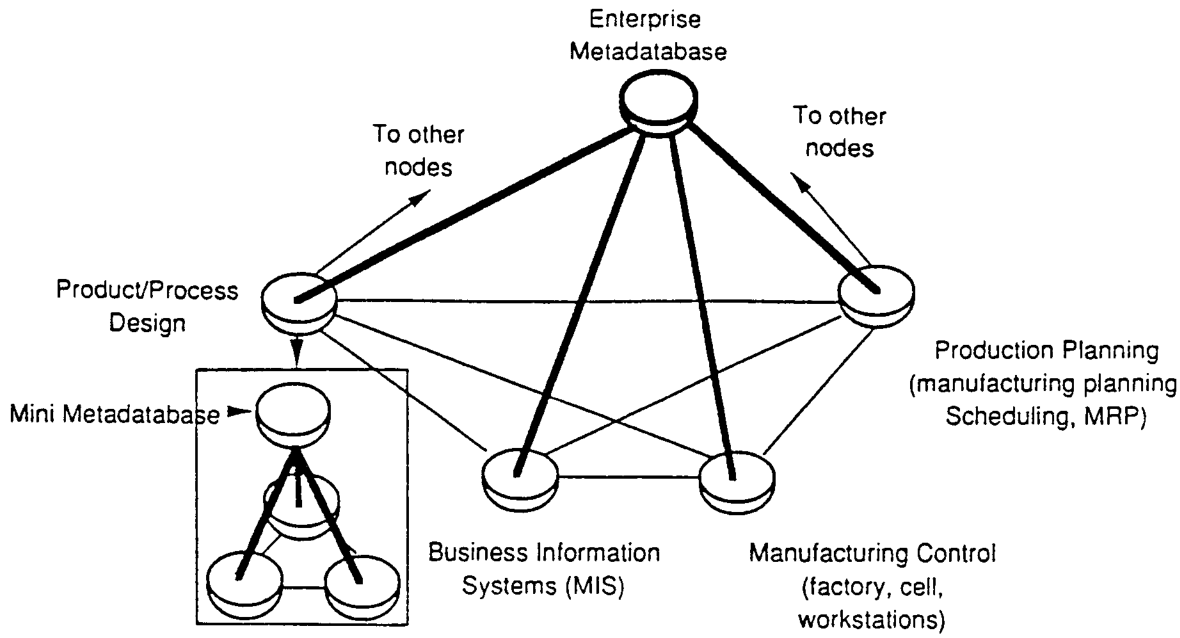

Figure 1. The CIM Architecture.

metadata, (2) a physical storage structure derived from this representation, and (3) a metadatabase management system providing facilities for effective implementation and processing of metadata. All these elements involve data and knowledge integration. Based on these capabilities, two more elements are required to fully support the envisioned functionality: (4) a global query construct using metadata as online knowledge to facilitate query formulation and processing for enterprise users and (5) a concurrent architecture managing distributed metadata (especially contextual knowledge) as the basis for achieving global synergy across functional (database) systems.

Several basic results have been achieved and successfully implemented in a prototype [7], since the metadatabase project started in 1987. They encompass all five elements but are nonetheless limited currently to manufacturing systems per se-- i.e., focusing mainly on shop floor control process design, MRP II and order entry. These results include:

(1). A general metadata model. This logical model, called Global Information Resources Dictionary (GIRD) (reported in [6]), can be compared to the IRDS work due to NIST [2], except that it, unlike the latter, considers explicitly heterogeneous environments where the necessary scope of metadata includes both data models and contextual knowledge. The representation method of GIRD is based on a data and knowledge modeling approach, TSER [3, 7], which, in turn, provides a ready capability for acquiring enterprise metadata through information modeling in heterogeneous environments.

(2). A rulebase method. The GIRD model includes as an integral part of it a rulebase submodel that consolidates contextual knowledge and combines it with the representation of data models in GIRD. An attendant rulebase processor was also developed to fire these production rules as well as to manage the rulebase.

(3). An architecture for metadatabase management system. Four basic categories of metadata storage and processing requirements were identified for the GIRD model: relational capabilities, semantic hierarchies, rulebase, and (complex) files (including possibly text, images, and routines). As a first step towards implementing the concept, the prototype metadatabase employs a relational platform (i.e., DEC Rdb) to couple with the rulebase processor. Semantic hierarchies and file management are achieved through both custom developed shells and standard utilities provided by the DEC micro-computing environment.
The custom shells also interface with users and employ the platform for transaction management.

(4). Global query formulation and processing. The metadatabase is envisioned and used as online knowledge to provide a direct method for query formulation. Enterprise (end) users articulate their queries with the global information model directly, as opposed to relying on any global languages and

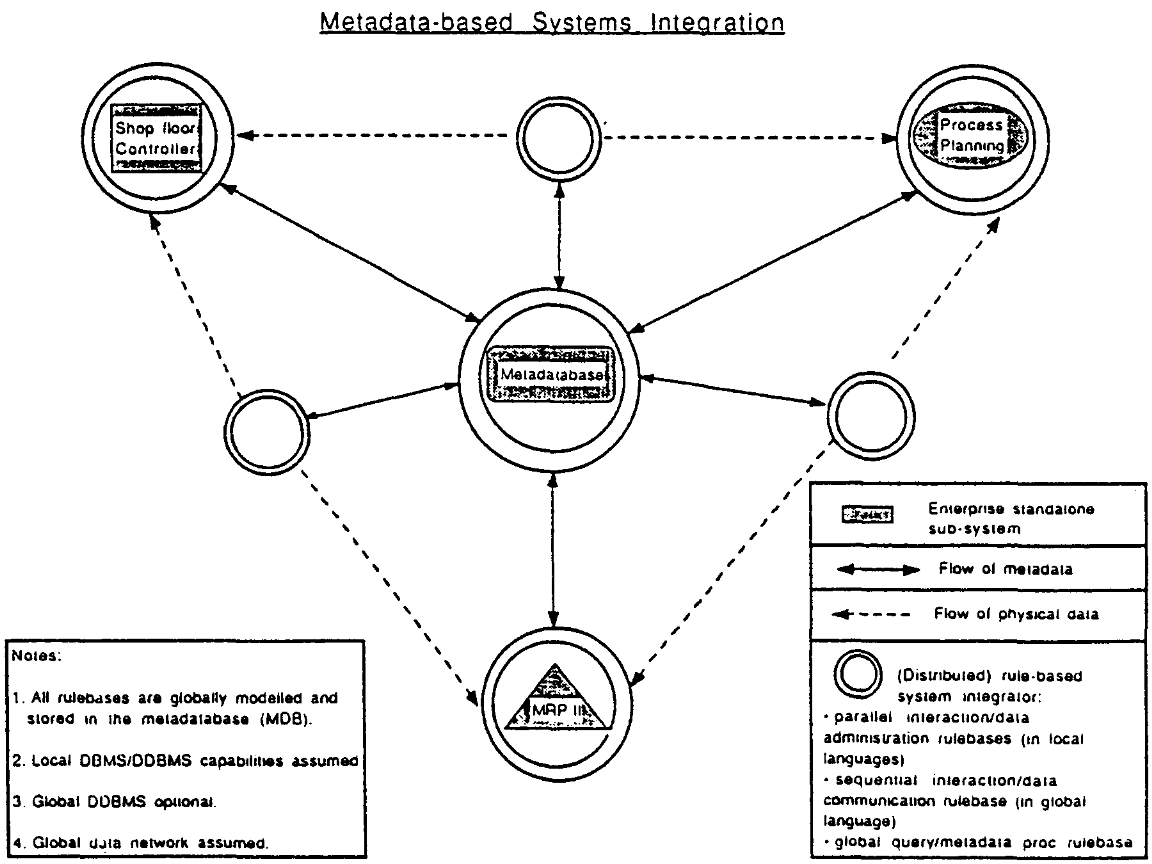

Figure 2. The Metadatabased Systems Integrator.

schemata integration. The query is then decomposed and coded into local data languages for processing. Metadata are used throughout this process to assist query formulation, optimization, and processing.

(5). The conceptual design of concurrent architecture. A central objective of the metadatabase model is to simplify the requirements of global synchronization by virtue of the use of metadata. Towards this end, a concurrent architecture was formulated and characterized in terms of distributed metadata capabilities, especially a distributed rulebase shell system as shown Figure 2, page 4.

This concept of metadata-based integration of concurrent systems requires all of the capabilities developed above and new major results proposed herein. Therefore, we now turn to the new plans where this concept is further discussed.

3. The Plan.

Reaping benefits from current capabilities for enterprise information integration, the metadatabase model provides a needed basis while requires further development. Three basic objectives will guide this effort:

(1). Extending the results to include engineering design functions and standards. The current capabilities need to be extended and verified for applications in the domain of engineering as well as manufacturing systems. In particular, the GIRD model and the TSER modeling method both need to consider the specific characteristics of certain product data models, especially select object-oriented constructs and standards such as PDES [14]. Emerging concepts for national and international standards will also be considered; a prime case is the CIMOSA work due to ESPRIT [1]. This principle is both a given and a strategy to deal with information integration.

(2). Pursuing concurrent operation of all functional systems with minimal or no global serialization on data processing. This principle invokes concepts such as federated or autonomous systems that have been employed for the management of multiple databases in the past decade. However, it goes beyond these approaches by explicitly calling for parallelism as an alternative to the von Neumann model of synchronization they tend to employ at a global level. The rationale for this strategy is two-fold: First, physically consolidating all existing functional systems into a super, single, all-encompassing system is not practical in many situations, due to business constraints (e.g., financial and organizational considerations) as well as the ever-evolving nature of technology that is destined to render any single system design dissatisfactory in a short period of time. Second, the computing complexity imposed on global synchronization in the envisioned environments tends to be intractable for non-trivial systems. Concurrent systems design, therefore, opens up new possibilities for simplification of the complexity.

(3). Converting the global control or "synchronization" problem in integration from one that deals directly with raw data (processing) to one focusing on metadata (management). This transformation of problem space is actually implied by the notion of "integration through information" in the above-stated objective. Specifically, we acknowledge the fact that integration is ultimately characterized in terms of the logical unity or functional synergies of the enterprise; which, in turn, can be represented through an enterprise information model comprising both data and knowledge abstractions. Moreover, we also recognize that all data synchronization efforts are based on some control logic and, hence, can be formulated as a class of knowledge and represented as such. Therefore, when both the information model and the control logic are formulated into a unified (online) kernel for the enterprise, a metadatabase results; which at least represents, and promises to facilitate or even implement, the manifested logic of integration. As such, the metadatabase system is fundamentally different from any kind of distributed database management systems that seek to synchronize data processings (through serialization) among multiple databases. Instead, it could be considered as a distributed rulebase management system that synchronizes pertinent metadata throughout the concurrent systems. The metadatabase itself maybe either centralized or distributed (e.g., either with multiple copies of it or a (recursive) hierarchy of main-metadatabase and mini-metadatabases for functional subsystems). The basic reason for this conversion of problem from raw data to metadata is this simple fact: metadata are more stable than raw data. In fact, one can expect that the number of "write" transactions an integrated enterprise would inflict on metadata be some order of magnitude less than that on raw data.

Together, these principles lead to three progressive levels of simplification for the complexity of the integration problem. First, just like all traditional federated or autonomous designs, the concurrent approach allows the possibility of simplification designs through what might be called "management by exception"; i.e., reduce the scope of global control to only those core information requirements that cut cross multiple subsystems while leaving other local, routine jobs to individual local information management systems. Second, with the concurrent approach, this strategy of management by exception can be maximized through capitalizing on the promises of parallelism enabled or facilitated through the metadatabase, thereby further reducing the need for global synchronization. Third, since the control strategy and the concurrent architecture are based mainly on metadata management and processing, their implementation can reap the benefits of the stability of metadata (vis-a-vis high volume transactions associated with raw data) as a fundamental simplification of complexity.

To strive towards these objectives from current results, the following major tasks are required.

(1). Develop a sufficient metadata model for both engineering and manufacturing systems. Specifically, a particular set of object-oriented capabilities will be identified for the proposed Express constructs of PDES [12], then the GIRD model and TSER constructs will be extended for this particular target so as to suffice the need of global information model in the extended enterprise. In a similar way, mappings and other forms of logical linkage between the metadata models and IRDS and CIMOSA will also be investigated.

(2). Develop a dedicated metadatabase management system integrating the four categories of metadata: relation, object, rule, and file. This work goes beyond the current prototyping effort, which was focused on illustrating the concept as opposed to providing the new software technology and architecture implied by the metadata storage and processing methods. Actually, this task in its own right amounts to an investigation of the integrated data and knowledge environment specified above.

(3). Develop a distributed rulebase technology implementing the concurrent architecture. This technology will supplement the current rulebase capabilities to manage and process the local contextual knowledge through the metadatabase according to the concurrent architecture. That is, together, the metadatabase will support the implementation of metadata to (a) create a consolidated rulebase model, (b) distribute pertinent context knowledge to local systems for parallelism, and (c) maintain the entire distributed metadata environment.

References

1. ESPRIT Consortium AMICE (eds.), Open System Architecture for CIM, Springer-Verlag, Berlin, Germany, 1989.

2. Goldfine, A. and P. Konig, A Technical Overview of the Information Resource Dictionary System, 2nd ed., NBSIR88 - 3700, National Institute of Standards and Technology, 1988.

3. Hsu, C., "Structured Database Systems Analysis and Design Through Two-Stage Entity-Relationship Approach," Proceedings 4th International Conference on Entity-Relationship Approach, IEEE Computer Society Press, pp. 56-63, 1985.

4. Hsu, C. and C. Skevington, "Integration of Data and Knowledge in Manufacturing Enterprises: A Conceptual Framework," J. of Manufacturing Systems, Vol. 6, No. 4, pp. 277-285, 1987.

5. Hsu, C. and L. Rattner, "Information Modeling for Computerized Manufacturing," IEEE Transactions on Systems, Man, and Cybernetics, Vol. 20, No. 4, pp. 759-776, 1990.

6. Hsu, C., M. Bouziane, L. Rattner and L. Yee, "Information Resources

Management in Heterogeneous, Distributed Environments: A Metadatabase Approach", IEEE Transactions on Software Engineering, June 1991.

7. Hsu, C., M. Bouziane, W. Cheung, J. Nogues, L. Rattner, and L. Yee, "A Metadata System for Information Modeling and Integration," Proceedings 1st International Conference on Systems Integration, IEEE Computer Society Press, pp. 616-624, 1990 (An extended version is under review for J. of Systems Integration).

8. Krishnamurty, V., S. Su, H. Lam, M. Mitchell, and E. Barkmeyer, "IMDAS, An Integrated Manufacturing Data Administration System," J. of Data and Knowledge Engineering, Vol. 3, No. 2, pp. 109-131, 1988.

9. Madnick, S., M. Siegel, and R. Wang, "The Composite Information Systems Laboratory (CISL) Project at MIT," Data Engineering Bulletin, Fall 1990, pp. 10-15.

10. McLeod, D., "Interbase: An Approach to Controlled Sharing Among Autonomous, Heterogeneous Database Systems," Data Engineering Bulletin, Fall 1990, pp. 4-9.

11. Rattner, L., Information Requirements for Integrated Manufacturing Planning and Control: A Theoretical Model, unpublished Ph.D. Thesis, Department of Decision Sciences and Engineering Systems, Rensselaer Polytechnic Institute, Troy, NY 12180-3590, December 1990.

12. Schenck, D., "EXPRESS Language Reference Manual," ISO TC184/SC4/WG1, Document N466, 1990.

13. Sheth, A. and J. Larson, "Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases," ACM Computing Surveys, Vol. 22, No. 3, pp. 103-236, 1990.

14. Smith, B., "Product Data Exchange Specification: First Working Draft," NISTIR 88-4004,NIST Order Number PB89-144794.

15. Thomas, G., G. Thompson, C-W. Chung, E. Barkmeyer, F. Carter, M. Templeton, S. Fox and B. Hartman, "Heterogeneous Distributed Database Systems for Production Use," ACM Computing Surveys, Vol. 22, No. 3, pp. 237-266, 1990.

16. U.S. Air Force, Integrated Information Support System (IISS) Report, Materials Laboratory, Air Force Systems Command, Wright-Patterson Air Force Base, Ohio, 1983.