Data Management Research at The MITRE Corporation

Arnon Rosenthal[1], Len Seligman[2], Catherine McCollum[2], Barbara Blaustein[2],
Bhavani Thuraisingham[1], Edward Lafferty[1]

[1] The MITRE Corporation, 202 Burlington Road, Bedford, MA 01730, USA
[2] The MITRE Corporation, 7525 Colshire Drive, McLean, VA 22102, USA

1. INTRODUCTION

The MITRE Corporation provides technical assistance, system engineering, and acquisition support to large organizations, especially U.S. Government agencies. We help our customers to plan complex systems based on emerging technologies, and to implement systems based on commercial-off-the-shelf products. In MITRE's research program, instead of emphasizing concerns of DBMS or CASE vendors, our research emphasizes the issues of organizations who need to use such products. For example, we favor areas where we can build over commercial products, rather than changing their internals.

Data management at MITRE goes beyond research, to include technology transition, system engineering, product evaluation, prototypes, tutorials, advice on customers' strategic directions, and participation in standards efforts. We use prototyping to illustrate potential improvements in customer systems, to understand vendors' capabilities, or both. There are close connections with efforts in object management, real-time systems, reengineering, artificial intelligence, and security.

This paper emphasizes the research efforts, grouped into five major themes: information integration, security and privacy, active and responsive systems, metrics, and digital libraries. For each theme, we list the major questions being explored, and identify projects and contacts for further information.

2. INFORMATION INTEGRATION

To integrate information across heterogeneous sources, one must (1) resolve infrastructure heterogeneity (e.g., different data models, DBMS products, etc.), (2) resolve representation and semantic heterogeneity in the data, and (3) perform instance identification and merging. Commercial products (e.g., gateways, replication servers, distributed query processors) are progressing rapidly in addressing infrastructure heterogeneity, so in that area MITRE emphasizes technology awareness and transfer, and assessment of customer systems (section 5). MITRE has concentrated most of its information integration research in addressing data heterogeneity, as discussed below. In instance reconciliation, we focus on knowledge-based techniques for special domains (e.g., target sightings).

Our customers integrate data to meet operational goals, in small, concurrent projects rather than combining entire databases. They use multiple modes of integration, such as view tables (relations or object classes that derive data from multiple systems), electronic data interchange (EDI) messages, and warehouses. While each mode requires somewhat different mediation algorithms, the knowledge required to drive mediation is mostly the same. We therefore seek designs for both mediators and knowledge bases that maximize reusability [RoSe94]. The potential savings are huge. For example, one system of systems anticipates spending several million dollars for custom-written data transfer programs.

Schema-level semantic integration is often the factor that limits the utility of distributed DBMSs. The solutions require knowledge of the problem domain and the legacy data, so vendor software will play a smaller role than in DBMS heterogeneity. Our specific research follows two major threads: *mediation* and *knowledge capture*.

Mediation: We are building a new kind of *semantic gateway* that allows clients to use their own data semantics (both schema and conventions for interpreting attributes) when accessing foreign databases. As a first step toward integration, this form of mediation seems low risk: most processing is local, the mediated query is executed by the data source without a distributed DBMS, and instance reconciliation is not required. It advances over earlier work (e.g., MIT's "source-receiver" model [SSR94]) by formalizing the semantics, allowing a receiver table to be derived from multiple source relations, and handling compound attributes (e.g., <Latitude, Longitude> versus Position).

(contacts: Arnon Rosenthal, James Scarano {arnie,jgs}@mitre.org)

Addressing data semantics outside databases, we extend the approach to distributed objects. Specifically, we provide a semantic gateway for arguments passed by a CORBA-compliant Object Request Broker [RoSc95] (contact: Myra Jean Prelle, mjp@mitre.org).

Knowledge capture and management: We are prototyping a repository to manage domain knowledge for integration, and several tools to populate it. For several small areas of Air Force operations we are developing conceptual schemas; where such schemas already exist, we are importing them. These *reference schemas* are used as hubs to which we relate data definitions in component systems. To permit scale-up and local autonomy, these hubs need to be independently extensible; at the same time, it will be necessary to interrelate and share portions of them (e.g., Measurement_Units).

To aid in designing the hubs and correlating them with existing databases, we have developed tools that identify candidate matches between attributes in different schemas. After all available descriptions (formal and informal) of each attribute are converted to free text, an information retrieval product (Personal Librarian) estimates similarity and forms clusters, which are then reviewed by the human expert. (contact: Ed Housman, emh@mitre.org).

We envision that a broad array of capture techniques will become available. However, analysis costs will remain high because the capture tools will never detect all matches, and humans will need to review candidate matches. Hence we design the knowledge base to expand reuse of assertions (both positive and negative). For example, large assertions (e.g., view definitions) may be too large to examine, edit, and reuse, so tools should capture small, modular assertions and derive the larger ones automatically. Also, the structures designed to make the knowledge base manageable (uncertainty about candidate assertions, multiple hubs, sharable partial specifications) are likely to evolve, and hence must be hidden from mediators. (contacts: Len Seligman, Arnon Rosenthal {seligman, arnie}@mitre.org).

Other Projects: In addition to the work mentioned above, we have built prototypes that demonstrate the utility of information integration in two major application areas: Command and Control (contact: Ray Emami, gemami@mitre.org) and Genomic Information Systems (contact: P-Y. Ryberg, pryberg@mitre.org). We also have begun explorations comparing fine- versus coarse-grained objects to encapsulate relational databases (contact: Chris Reedy, creedy@mitre.org) and examining configuration and evolution issues in federated schemas (contact: Chris Bosch, cbosch@mitre.org).

3. INFORMATION SECURITY AND PRIVACY

MITRE has an active research program addressing security and privacy in DBMSs. This involves both the incorporation of multilevel security models primarily used to protect classified data (e.g., Secret, Top Secret, etc.) and the development of models and mechanisms for protecting the privacy and integrity of other types of sensitive data (e.g., proprietary, financial, and sensitive personnel data). MITRE defined and prototyped some of the earliest multilevel secure (MLS) database systems; today, through a combination of contract and internally funded research, MITRE is working in several key areas of MLS database management and database privacy: secure distributed data management, secure object-oriented data management, integrity protection and dynamic separation of duty controls, database inference control, and multilevel DBMS performance assessment.

The Secure Federated Data Management project (contact: Catherine McCollum, mccollum@mitre.org) is concerned with allowing cooperating organizations to share sensitive data across jurisdictional boundaries in a federation, even when organizations have different confidentiality and handling requirements. The goal of the effort is to provide a foundation for data interchange controls that cope flexibly with diverse security needs and respect the autonomous authority of each organization. Data interchange agreements protect sensitive data by tying access privileges to responsibilities to safeguard the data accessed. Cooperating parties to an agreement retain their autonomy in general but, through the agreements, concede part of it specifically with respect to particular data items in exchange for desired access to information belonging to the other organization. Agreements are expressed in a specification language and enforced automatically. We have developed a prototype, running on a commercial DBMS, that examines database requests and enforces the agreed-upon handling procedures, and a prototype tool for defining agreements [BMRS95].

The MUSET MLS Distributed DBMS project (contact: Barbara Blaustein, btb@mitre.org), is concerned with algorithms for secure execution of multilevel transactions. These transactions include operations that write data at multiple security levels; current MLS DBMSs constrain operations at a single level to be executed within user sessions at that security level. We have developed a model,

algorithms, and a policy for interpreting atomicity in a multilevel security context; analyzing the degree of multilevel atomicity that can be achieved for a transaction; and breaking up transactions into sections that can be reordered, within the constraints of the transaction's read-write dependencies, to execute without illegal information flow [SBJN96]. We have also analyzed the use of two-phase commit transaction management techniques for the execution of distributed single-level transactions in a multilevel secure environment.

The Multilevel Secure Object-Oriented DBMS (MLS OODBMS) project (contact: William Herndon, wherndon@mitre.org) is developing a research prototype of an MLS OODBMS model and studying architectural approaches and assurance issues for MLS OODBMSs. We are investigating MLS support for important OODBMS features that are missing from earlier MLS OODBMS models. These include collections, fine-grained labeling, and class hierarchies that are free of monotonicity restrictions. To date we have developed and explored the Unrestricted Fine-grained Object Security (UFOS) model, an MLS object-oriented data model designed to be compatible with the commercial OODBMSs. We have also studied architectural issues for supporting the UFOS model within typical persistent object and object-relational systems and are currently prototyping the UFOS model through modification of TI's Open OODB research prototype [RHWT94].

The Inference Project (contacts: Bhavani Thuraisingham, Marie Collins, {thura, mcollins} @mitre.org) addresses the problem of database users inferring information which they are not authorized to acquire. Existing MLS DBMSs only allow users access to data for which they are authorized, but they cannot control inferences which result from users issuing multiple (authorized) requests. Based on differences in types of inferences, this research has pursued two approaches. One approach is to process security constraints (rules that assign security levels to the data) during query, update, and database design operations. We have developed proof-of-concept prototypes for processing these constraints. The second approach is to use knowledge-based techniques to develop inference controllers which would act as advisors to the systems security officer. While security constraints provide an effective mechanism to handle limited forms of inferences, we believe that knowledge-based inference control techniques are necessary to handle a broad range of inference types. We have also developed a knowledge-based inference controller tool [ThFo95].

Other recent research efforts in the area of security include a DBMS integrity model, an MLS DBMS benchmarking methodology, and an MLS database design tool. The DBMS integrity research focus is the integration of additional integrity and dynamic separation of duty models into a relational DBMS [NBM95] in order to enforce application-level integrity. The MLS DBMS benchmarking methodology extends DeWitt's Wisconsin benchmark with data attributes and experiments to measure the impact of variations in the MLS DBMS architecture, database (the amount of data at different classification levels), and workload (the levels at which database operations originate). A benchmarking tool has been developed that implements this methodology, and a number of MLS DBMS products have been studied [DHJM94]. The MLS database design tool supports database designers in considering trade-offs between confidentiality and data integrity for commercial trusted MLS DBMS products.

4. ACTIVE AND RESPONSIVE SYSTEMS

Much research in this area has focused on the internals of DBMSs that support active behavior (e.g., semantics and efficient implementation of Event-Condition-Action rules). MITRE's research has focused more on the ways in which active and responsive databases can be exploited in order to meet the requirements of our clients' applications. We are exploring the use of these systems for maintaining approximately consistent materialized views, for supporting real-time command and control applications, for doing situation monitoring in multidatabase federations, and for supporting security and integrity policy enforcement throughout a workflow.

Quasi-views: (contact: Len Seligman, seligman@mitre.org). Many automated information systems need to: (1) transform and cache information from dynamic, shared databases, (2) reason about the current state of those data, and (3) perform long-running tasks but cannot lock the objects about which they are reasoning, so as to allow concurrent access by other applications. Many of these applications can tolerate some deviation between the state of their caches and that of the shared databases, as long as this deviation is within specified tolerances.

We present a new approach to maintaining approximate consistency of client caches for such applications [SeKe93]. We have introduced and formalized *quasi-views*—materialized, object-based views defined over shared databases that are refreshed according to staleness conditions and refresh strategies (e.g., eager, lazy) specified

declaratively by application designers. A novel layer of software called a *Mediator for Approximate Consistency* automatically generates rules and other database objects that fulfill stated consistency requirements, shielding the application developer from implementation details. We have implemented a prototype and have extended the approach to support both active and passive (e.g., legacy) data sources.

Real-time Data Management for Command and Control Applications: (contacts: Bhavani Thuraisingham, Edward Bensley, {thura, ebensley} @mitre.org). We have defined requirements and are prototyping a real-time system infrastructure for command and control [Bens95, ThSc94]. The goal is a modular, open, distributed architecture based largely on commercial operating system, data management, and networking components.

For the data management component of the infrastructure, we propose a real-time data manager consisting of a shared memory data manager for real-time transaction processing and a persistent storage data manager for ad hoc query processing. We are implementing the shared memory data manager ourselves, and are using ZIP_RTDBMS, a commercial real-time DBMS product over the Lynx real-time operating system, for the persistent data. In addition, we have implemented an Event-Condition-Action rule manager on top of ZIP_RTDBMS. We are addressing transaction scheduling issues, which present a difficult challenge. In addition to serializability and meeting time constraints, one must satisfy transaction precedence and requirements for recent data.

Other Projects: We are developing techniques for efficient situation monitoring in a multidatabase federation (contact: Patricia Carbone, carbone@mitre.org). Users define alerters (i.e., rules whose action is to alert a user or application) in terms of an import schema of a federated database. These alerters are transformed into "agents" which monitor component data sources and those which evaluate conditions that span multiple data sources. We have defined three alternative architectures for doing situation monitoring in multidatabase environments and are analyzing their performance implications. We are developing a prototype implementation of one of these approaches, which leverages commercial "asynchronous replication" products as well as active database rules. Future plans include the automatic generation of these alerters. The approach is to do sensitivity analysis on a Bayesian inference network (which represents a plan for accomplishing some mission) to determine what data to monitor.

MITRE is also pursuing the use of active and responsive systems for managing sensitive data. [Smit94] investigates a new rule execution algorithm for MLS rules. MLS security constraints are preserved as well as arbitrary user-defined execution orderings by using an adaptation of the standard multiversion concurrency control algorithm. Another effort (contact: Barbara Blaustein, btb@mitre.org) is underway to tie security and integrity policy enforcement to data exchanged as part of an inter-organizational workflow.

5. METRICS AND ASSESSMENT

To manage their portfolios of data-intensive applications, organizations must make strategic decisions (e.g., which systems to retire, which ones to integrate). At the same time, they wish to improve their organizational process for managing data. MITRE has developed complementary assessment techniques to evaluate the sustainability of an implemented database system and the maturity of an organization's data management processes. We have begun developing analogous metrics for information retrieval from heterogeneous sources.

Strategic decisions (e.g. to extend, port, or retire a system) require an understanding of the difficulty of a task. For data-intensive systems, one can be guided by factors such as consistent use of standards, error control, self-descriptiveness, and application independence from low-level data structures. To evaluate these systems with speed and objectivity, tools and methodologies are needed. The Metadata Assessment and Risk Summary (contact Pamela Campbell, pcampbel@mitre.org) provides an on-line questionnaire which leads a skilled database analyst through a detailed method for assessing databases of the most common types: file-based, hierarchical, network, and relational. It also automatically identifies some questionable constructs (e.g., use of system-dependent filenames), subject to an analyst's judgment. The results include an objective, repeatable scoring of many factors that affect how easy it would be to sustain a system. The assessment also examines the ability of a system to prevent data corruption (e.g., support for referential integrity), to be tuned for performance (e.g., availability and use of indexes and clustering), and to use industry standards (e.g., for data access and data transfer). Next, one maps these factors into metrics (e.g., use of a 4GL hurts portability but helps maintenance). The final report translates the findings into risk drivers and risk mitigators for maintenance, evolution, and porting.

Complimentary to the assessment of existing applications, the Data Management Maturity Model and Assessment Tool (contact: Burton Parker, bgparker@mitre.org) assess the *organizational* activities across the data life cycle [PSS94]. This process-oriented evaluation adapts and extends the Software Engineering Institute Capability Maturity Model to enterprise data management. It considers activities necessary for individual database development projects (e.g., configuration management of data models and database design) as well as those necessary to provide effective data sharing across business functions and multiple information systems (e.g., quality assurance review for compliance with standardized data models, data elements, and business rules). The existence of these (and many more) activities contribute to determining the overall maturity level of the organization and to identifying critical areas of improvement.

The Metrics for Integrated Data Access project (contacts: Len Seligman, Marcia Kerchner, {seligman, kerchner}@mitre.org) is investigating metrics for technologies that support information retrieval across heterogeneous data sources (e.g., text, imagery, structured data, etc.). We have found some widely used retrieval metrics (e.g., precision and recall) to be of limited value for interactive, iterative searches that span heterogeneous sources. In addition, many existing metrics do not assess an approach's support for the end-to-end retrieval process, including query formulation, result merging, and information presentation. We are now developing customer user scenarios and populating databases that will be used for empirical evaluations and have initiated discussion of metrics within the research community [Seli95].

6. DIGITAL LIBRARIES

MITRE has several research projects in digital libraries, addressing information tagging, metadata management, information organization, and tools for information retrieval. MITRE is leading a Coalition for Networked Information (CNI) initiative to explore architectures, standards, technical and information management issues for advancing networked information discovery and retrieval (NIDR). The Digital Library Metadata (DLM) project is developing an experimental metadata framework that enables integration of application-specific metadata with the Z39.50 standard. The Image Tagging for Digital Libraries project has developed an Imagery Generalized Markup Language (IGML) to allow users to retrieve images based on feature level tags. The Electronic Libraries and Visual Information Study (ELVIS) is researching the improvement of information retrieval through query expansion using terms gathered from thesauri and added to a baseline query. The Smart Yellow Pages project is developing techniques for categorizing and disseminating text documents in a dynamic, distributed environment. Finally, we have developed a MITRE Information Infrastructure (MII) for cross-platform publishing, browsing, search, dissemination, and intelligent push-pull of digital information. The MII serves as both a useful, operational system and as a testbed for empirical research. For further information on MITRE digital libraries research contact Ray D'Amore (rdamore@mitre.org) or Michael Josephs (mrj@mitre.org).

REFERENCES

A more complete reference list and several of the cited papers are available at **`http://www.mitre.org/pubs/data_mgt`**.

Information Integration

[RoSc95] A. Rosenthal and E. Sciore, "Description, Conversion, and Planning for Semantic Interoperability," *IFIP WG6.2 Conference on Data Semantics*, Atlanta GA, 1995.

[RoSe94] A. Rosenthal, L. Seligman, "Data Integration in the Large: The Challenge of Reuse," *Int. Conf. on Very Large Data Bases*, Santiago Chile, Sept. 1994.

[SSR94] E. Sciore, M. Siegel, A. Rosenthal, "Using Semantic Values to Facilitate Interoperability among Heterogeneous Information Systems," *ACM Transactions on Database Systems*, June 1994.

Security and Privacy

[BMRS95] B. Blaustein, C. McCollum, A. Rosenthal, K. Smith, and L. Notargiacomo, "Autonomy and Confidentiality: Secure Federated Data Management," *2nd Intl. Conference on Next Generation Information Technologies and Systems*, Naharia, Israel, June 1995.

[DHJM94] V. M. Doshi, W. Herndon, S. Jajodia, and C. McCollum, "Benchmarking Multilevel Secure Database Systems Using the MITRE Benchmark," *Procedings of the Tenth Annual Computer Security Applications Conference*, Orlando, FL, Dec. 1994.

[NBM95] L. Notargiacomo, B. Blaustein, and C. McCollum, "Merging Models: Integrity,

Dynamic Separation of Duty, and Trusted Data Management," *Journal of Computer Security,* to appear.

[RHWT94] A. Rosenthal, W. Herndon, J. Williams, and B. Thuraisingham, "A Fine-grained Access Control Model for Object-Oriented DBMSs," *Database Security, VIII Status and Prospects*, J. Biskup, et al. (eds.), North-Holland, Amsterdam, 1994.

[SBJN96] K. Smith, B. Blaustein, S. Jajodia, and L. Notargiacomo, "Correctness Criteria for Multilevel Transactions," *IEEE Transactions on Knowledge and Data Engineering (Special Issue on Secure DBMS Technology)*, to appear February 1996.

[ThFo95] B. Thuraisingham, and W. Ford, "Security Constraint Processing in a Multilevel Secure Distributed Environment," *IEEE Trans. on Knowledge and Data Engineering*, 7(2), 1995.

Active and Responsive Databases

[Bens95] E. Bensley et al., "Evolvable Systems Initiative for Real-time C3," to appear in the *Proceedings of the IEEE Complex Systems Conference*, Orlando, FL, 1995

[SeKe93] L. Seligman and L. Kerschberg, "An Active Database Approach to Consistency Management in Data- and Knowledge-based Systems," *Int. Journal of Intelligent and Cooperative Information Systems*, 2(2), 1993.

[Smit94] K. Smith, "Execution Reordering for Multilevel Secure Rules," *Proceedings of the Fourth Int. Workshop on Research Issues in Data Engineering: Active Database Systems (RIDE-ADS)*, Houston, TX, February, 1994.

[ThSc94] B. Thuraisingham and A. Schafer, "RT-OMT: A Real-time Object Modeling Technique for Designing Real-time Database Applications," *Proceedings of the 2nd Real-time System Applications Workshop*, Beltsville, MD, July 1994.

Metrics and Assessment

[PSS94] B. Parker, D. Smith, and D. Satterthwaite, "Data Management Capability Maturity Model," *Proceedings of the 11th DOD Database Colloquium*, San Diego, CA, 1994.

[Seli95] L. Seligman, "Metrics for Heterogeneous Data Access: Is there any hope?" Panel description in *Proceedings of VLDB*, Zurich, Switzerland, 1995.

ACKNOWLEDGMENTS

The authors would like to thank Pamela Campbell, Linda Chambless, and Manette Lazear for providing ideas, project descriptions, and valuable feedback. We would also like to thank Ron Haggarty, MITRE's Vice President for Research and Technology, for his committment to research in this area.