

The InfoSleuth Project: Intelligent Search Management via Semantic Agents

**Darrell Woelk and Christine Tomlinson
Microelectronics and Computer Technology Corporation (MCC)
3500 Balcones Center Dr.
Austin, Texas 78759**

Abstract

InfoSleuth is a research project at MCC that is developing and deploying new technologies for finding information available both in corporate networks and in external networks, such as networks based on the emerging National Information Infrastructure (NII). The InfoSleuth research is based on the Carnot technology that has been developed at MCC over the last 4 years. Carnot has been successfully used to integrate heterogeneous corporate information resources. Carnot executes queries in a distributed environment by dispatching autonomous computing agents to remote sites where they access databases and cooperate among themselves to properly merge resulting data into understandable information. The InfoSleuth project will investigate the use of Carnot technology in a more dynamically changing environment, such as the Internet, where new information sources are constantly being added and for which there is no formal control of the registration of new information sources. In this type of environment, traditional techniques for expressing and optimizing database queries are inadequate because of the rapidly changing schema information and the fuzzy nature of the queries. InfoSleuth will build on Carnot semantic modeling capabilities to enable "deep" descriptions of available information sources. InfoSleuth will deploy semantic agents to carry out distributed, coordinated, self-adapting search algorithms. MCC is seeking industrial sponsors to participate in the InfoSleuth research.

1. Introduction

InfoSleuth is a consortial research project at MCC to develop and deploy new technologies for finding information available both in corporate networks and in external networks, such as networks based on the emerging National Information Infrastructure (NII), for example, MCC's EINET, [EINET].

It is focusing on the problems of locating, evaluating, retrieving, and merging information in an environment in which not only new information, but more importantly, new information sources are constantly being added and for which there is local autonomy, i.e., no network-wide control of the registration of new information sources and their content.

In this type of environment, traditional techniques for expressing and optimizing database queries are inadequate because of the rapidly changing schema information and the fuzzy nature of the queries. Text search techniques and interactive navigation techniques are also inadequate because of the immense size and (potentially) remote distribution of the available information.

InfoSleuth will build on the MCC Carnot technology, [WOEL93]. In particular, Carnot's semantic

modeling capabilities, [HUHN92] and [WOEL92], will enable “deep” descriptions of available information sources, and Carnot’s Extensible Services Switch, [TOML92], will enable InfoSleuth to deploy semantic agents to carry out distributed, coordinated, self-adapting search algorithms. An agent may either return information immediately and retire or may remain active in the network watching for information of interest to be added.

Section 2 of this paper will describe in more detail the motivations for the InfoSleuth project. Section 3 will describe the MCC Carnot technology upon which InfoSleuth is being built. Section 4 will describe the technical approach being used to develop the InfoSleuth software.

2. Motivations for InfoSleuth

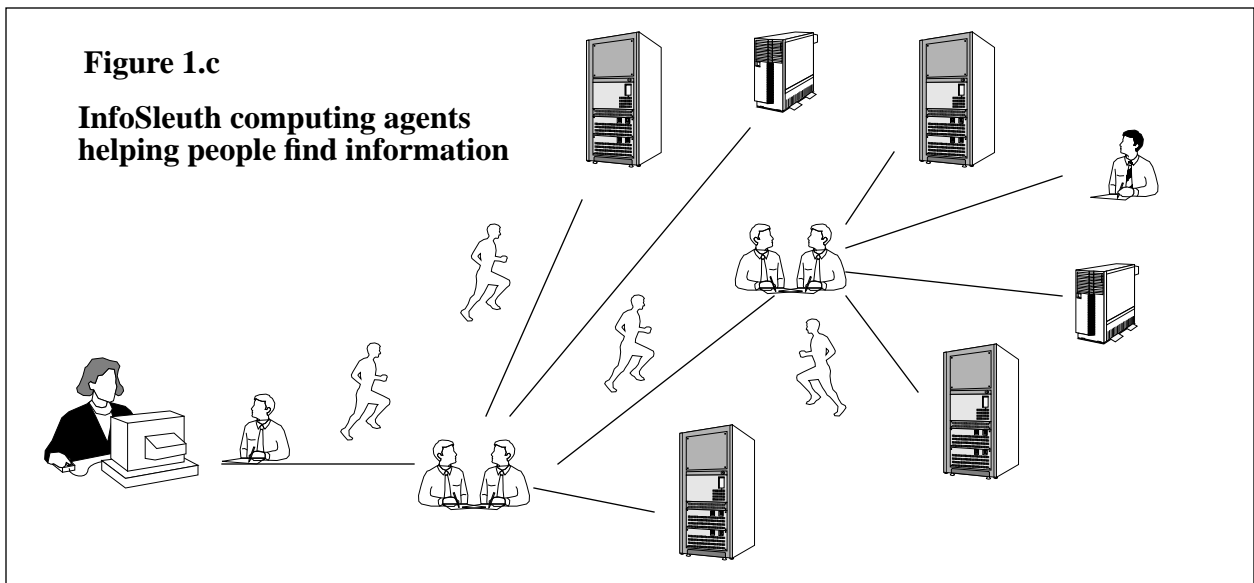
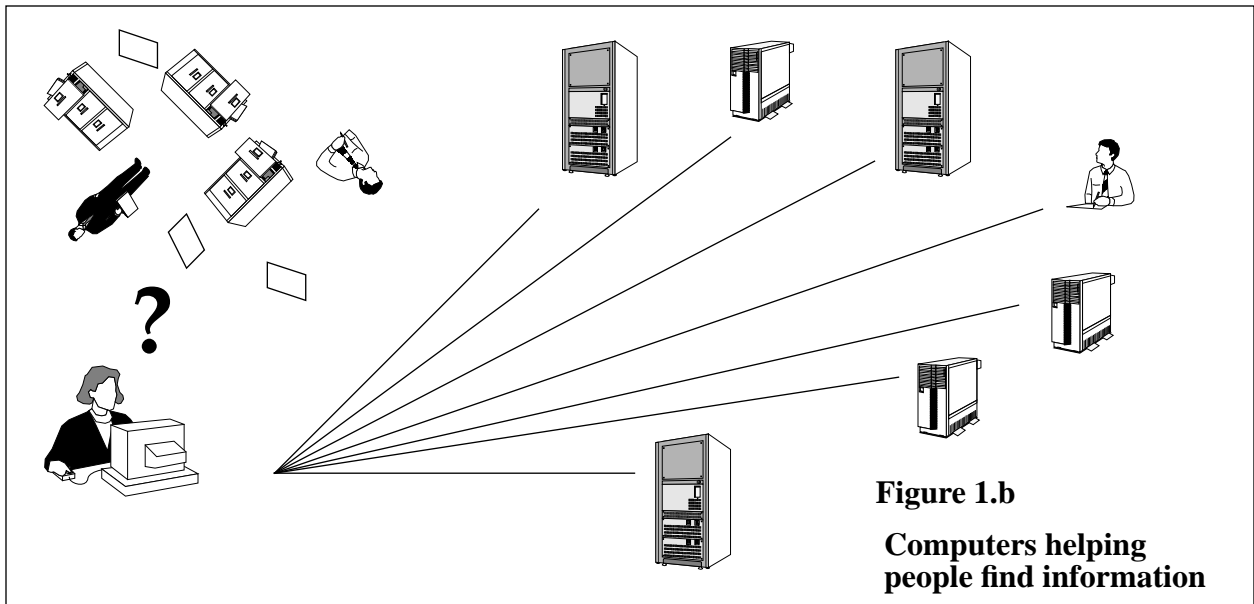
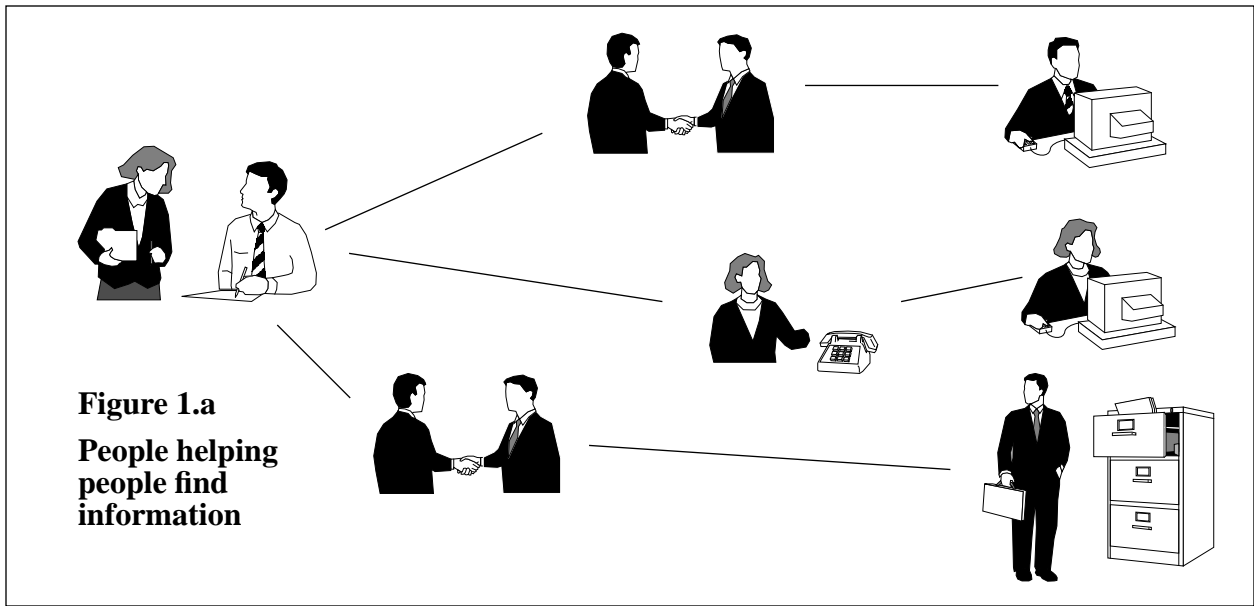
The number of different remote sources of information is increasing rapidly. Every organization depends on information for efficiently executing its business mission. Until now, two general categories of information were used in running an organization.

The first category is *Formal Information*, which is generated and used during the day to day operations of a business. A manufacturing company generates design documents, manufacturing schedules, quality assurance reports, inventory reports, sales forecasts, sales reports, budget forecasts, financial reports, etc. Application software generates this machine-readable information. The development of this application software requires a precise understanding of the information and the processes that are used to run the company.

Swirling around most businesses is information that falls into a second category, *Informal Information*. Informal Information includes the following:

- *Personal* databases, spreadsheets, and documents created by an individual to support the activities of the individual within the organization.
- *Transient* databases, spreadsheets, and documents generated within the company to document an activity that has not yet been identified as a candidate for becoming a source of Formal Information.
- *Informal* email interactions and working documents exchanged by individuals.
- *Inter-divisional* Formal Information that might be useful periodically to individuals in another division of the company if its existence were known.
- *External* information available via external public and private networks that might be useful to an individual if its existence were known.

Until a few years ago, most informal information was not in machine readable form. It was embedded in type-written documents, hand-written notes, reference books, and people’s minds. Figure 1.a illustrates the manner in which informal information was shared. A person would discover the existence of informal information through systematic or serendipitous interactions with other people. An example of systematic interaction would be a marketing manager requesting an analyst to determine the optimal geographical location for test marketing a new product. The marketing analyst might then access some formal information in the company concerning the product, but the analyst would also depend on interactions with others to find needed market information that might be unique to this product. It was not necessary for the analyst to know of the existence of formal information somewhere and how to physically access the information. The



analyst only had to know someone who knew how to find it.

The people in Figure 1.a served two purposes. First, each person had some specialized knowledge of an area, including knowledge of the specialties of other people. Second, two people could adapt their interactions to handle slight variations in the way a question is asked or to handle confusion on the meaning of information that is found.

As computers and communication among computers has become more ubiquitous, the information that the marketing analyst is looking for is more likely to be accessible directly. The need for all of the people in Figure 1.a seems to have been eliminated. In fact, it appears that the need for the marketing analyst has also been eliminated since the marketing manager should be able to directly access all of the information as shown in Figure 1.b. However, while the formal information stored in the remote computers is now available, the informal network of information represented by the people in Figure 1.a has been replaced in Figure 1.b by a web of static links and syntactic indices.

Figure 1.c illustrates the use of InfoSleuth computing agents to solve the problem. InfoSleuth will provide software tools that assist people in expressing their knowledge about information sources in a clear and concise manner. Using this knowledge, an InfoSleuth agent can potentially replace many of the human agents in Figure 1.a. An InfoSleuth agent can take the place of the marketing analyst. The marketing manager requests information from the agent through a natural interaction. The agent then independently sets out to find the information, possibly returning to the manager for clarification or to report status. The agent then interacts with other InfoSleuth agents to find the proper sources of information, to retrieve the information from those sources, and to present the information to the manager in an understandable manner. The marketing manager can request that the InfoSleuth marketing analyst agent continue to monitor information sources to identify changes in the requested information that might represent trends in the market.

The InfoSleuth marketing analyst agent might contact an InfoSleuth real estate investment agent to better understand changes in the real estate market in various geographical locations. The real estate investment agent may access information in a number of different online databases. Just as with a human agent, the agent will receive compensation for its assistance in addition to the compensation to the owner of the online database. A good real estate investment agent will have many customers and owners of databases will seek to have their databases accessed by that agent.

3. MCC Carnot Technology

The Carnot Project at MCC was initiated in 1990 with the goal of addressing the problem of logically unifying physically-distributed, enterprise-wide, heterogeneous information. Carnot provides a user with the means to navigate information efficiently and transparently, to update that information consistently, and to write applications easily for large, heterogeneous, distributed information systems. A prototype has been implemented that provides services for:

- enterprise modeling and model integration to create an enterprise-wide view,
- semantic expansion of queries on the view to queries on individual resources, and
- interresource consistency management.

Carnot also includes technology for 3D visualization of large information spaces, knowledge discovery in databases, and software application design recovery.

A key technical problem addressed by Carnot is the need to simplify the development of enterprise-wide applications that access information and keep information consistent. This requires that the Carnot system maintain a semantically rich understanding of the information used to run the enterprise. This understanding, in the form of a model of the enterprise, is kept in a knowledge base that is part of the Carnot system. Of course, the real data about the operation of the enterprise are maintained in various physical resources, such as databases, files systems, and application programs.

Once a model of the enterprise has been created, the database schemas can be individually related to this model. When this step is completed, each operation on an individual database schema has an equivalent operation on the enterprise model and an operation on the enterprise model can map into operations on multiple databases.

Furthermore, business rules that in the past were embodied in application programs can be represented in the enterprise model where the Carnot system can enforce them, thus simplifying the development of new application programs. Also, as new computer hardware, database management systems, or databases are added to the enterprise, they are also individually related to the enterprise model. The result is a powerful system that enables an unlinking of applications from physical resources. Applications do not need to change as a business expands. And applications do not need to change when two businesses merge.

The implementation of the Carnot system has required unique advances in two technology areas. First, innovative techniques for knowledge representation have been developed to capture and maintain an enterprise model and to map operations between an enterprise model and the physical databases. Creative new tools have also been developed to discover new knowledge in existing databases, code, and other artifacts. Second, a flexible, dynamic, distributed processing environment has been developed that supports the automatic generation of program scripts that execute on heterogeneous, distributed systems. The scripts control the flow of processing and can be reconfigured dynamically to respond to changes in the hardware environment or to the incorporation of additional information resources. The scripts are embedded in autonomous computing agents that can be dispatched to remote sites.

Carnot has developed and assembled a large set of generic facilities that are focused on the problem of managing integrated enterprise information. These facilities are organized as five sets of services as shown in Figure 2: communication services, support services, distribution services, semantic services, and access services. Figure 2 lists example services at each of the layers. A subset of these services have been implemented in each layer. The communication services provide the basic connectivity among hardware systems. The support services implement basic network-wide services, such as RDA, that are available to applications and other higher level services. The Extensible Services Switch (ESS), [TOML92], is software that supports the Carnot computing agents, [HUHN91] and [HUHN94]. The distribution services provide control of the interaction among multiple distributed systems. Examples of services at this layer are transaction management and workflow management. The semantic services provide an enterprise-wide view of all resources integrated within a Carnot-supported system. The Enterprise Modeling and Model Integration facility, [HUHN92], uses a large knowledge base as global context and a federation mechanism for integration of concepts from various heterogeneous models and databases.

The Carnot prototype software has been used by the sponsors of the Carnot project to develop a

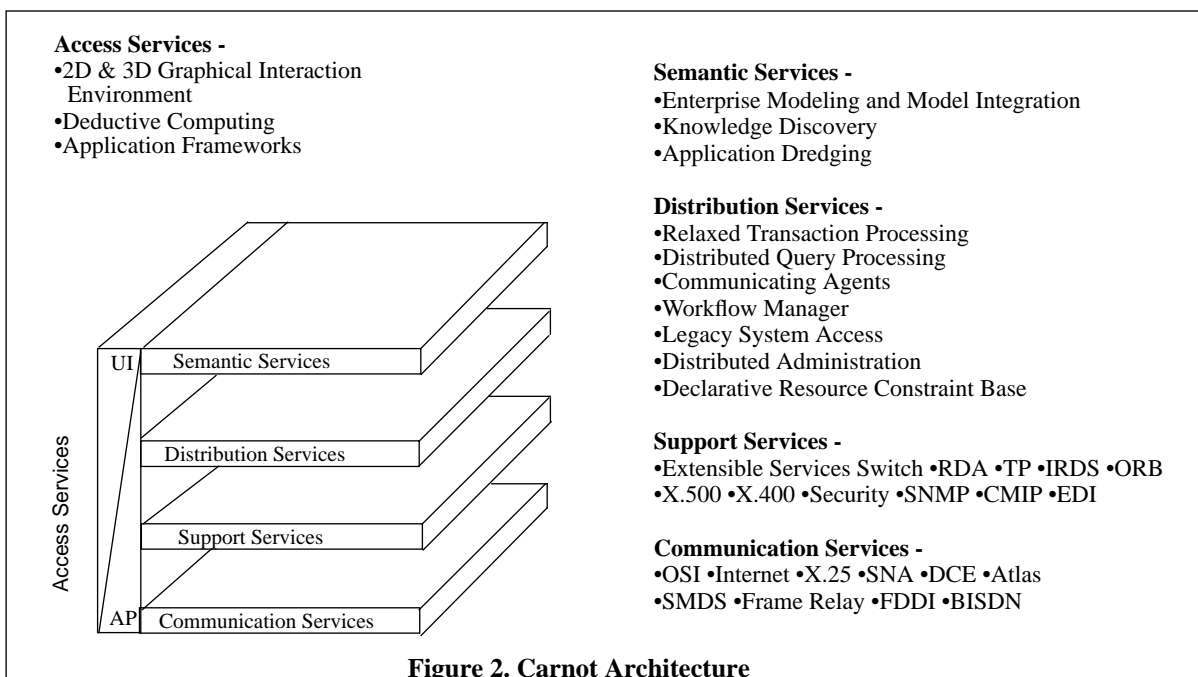
number of applications. These applications have included workflow management, heterogeneous database access, and knowledge discovery in large databases. Of particular pertinence to the InfoSleuth project is an application that integrated access to both a text database and a relational database from a single initial query.

4. The InfoSleuth Approach

The goal of the InfoSleuth Project is to develop and demonstrate technology that will expedite the process of searching for pertinent information in a geographically distributed and constantly growing network of information resources. In this type of environment, there can be no centralized database administrators or systems analysts to document database semantics or to write application programs that smooth over differences in access languages and database structures. InfoSleuth must be able to efficiently sift through the multitudes of potentially relevant information sources and discover the information that is pertinent to an individual or organization.

Such information is constantly being added to and deleted from the environment, and may include categories of information being added to an existing source or the addition of entirely new information sources. To avoid the problem of centralized administration of information, InfoSleuth will extend the notion of registration of services that has been successfully used in distributed environments, such as OSF Distributed Computing Environment (DCE) and OMG Common Object Request Broker Architecture (CORBA).

- Providers of information will *advertise* its availability. Information will be advertised in order to share information (or services) within a company or to generate revenue by selling information (or services).
- Clients that are searching for information will *discover* the availability of potentially useful advertised information. Autonomous InfoSleuth agents will be deployed to search for information, remaining active to monitor for the addition of pertinent new information.



- Clients will *fuse* information from many information sources. InfoSleuth agents will cooperate with each other to implement this fusion.

In the InfoSleuth environment, information is advertised by describing its information content in terms of a network-wide distributed taxonomy. This taxonomy is similar to a dictionary or directory but contains more information concerning the meaning of an entry and its relationship to another entries. This is an improvement over conventional *network yellow pages*, which do not really make it easier for clients to discover the availability of new information. The InfoSleuth taxonomy is like having the semantics of various yellow pages categories on-line.

Various interface tools will be developed to simplify the process of advertising the availability of information. These tools will utilize existing indexing techniques such as WAIS [WAIS], Gopher [GOPH], World Wide Web Worm [McBR94], and ALIWEB [KOST94] and other robots wherever possible. New tools will be developed that will concentrate on using natural language processing to assist an information provider in describing the information to InfoSleuth.

The addition of new categories and the refinement of the description of existing categories will lead to semantic ripples throughout the information space. As new information sources are advertised, agents that are looking for changes in the semantic background may *trigger* and report the presence of new or (apparently) more relevant information sources in the environment.

Once an information source has been advertised, it can be discovered by an InfoSleuth agent acting on behalf of a client. This discovery process is an extension of the analogous concept of having a client request a service in a distributed network. The discovery process differs from a request for a service or a query for information in that discovery is actually a process. After a set of InfoSleuth agents are given a description of the information to be found, they are dispatched into the network to search for meaningful information sources. At the request of the client, the InfoSleuth agents may remain active in the network watching for new information to be advertised.

Various interfaces will be used for requesting information from an InfoSleuth including a natural language interface that will be developed using the Knowledge Based Natural Language software developed at MCC [SING94].

Once information has been discovered by an InfoSleuth agent, the agent may travel to more than one information source and fuse the information from these information sources for presentation to a client. Multiple InfoSleuth agents may also cooperate to fuse information from multiple sources. The information fused by an InfoSleuth agent may itself become a new value-added information source that can be advertised and is available for discovery by other clients.

5. Status of the InfoSleuth Project

Initial InfoSleuth planning and prototyping has begun within the Carnot project at MCC. InfoSleuth will become a separate project beginning in January, 1995. MCC is seeking industry sponsors to participate in the InfoSleuth project.

References

- [GOPH] gopher://gopher.micro.umn.edu:70/11/Information%20About%20Gopher.
[EINET] http://galaxy.einet.net/galaxy.html.

- [HUHN91] Huhns, Michael N. and David M. Bridgeland, "Multiagent Truth Maintenance," IEEE Transactions on Systems, Man, and Cybernetics, Vol. 21, No. 6, November/December 1991, pp. 1437-1445.
- [HUHN92] Huhns, M., N. Jacobs, T. Ksiezzyk, W.M. Shen, M. Singh, and P. Cannata, 1992."Enterprise Information Modeling and Model Integration in Carnot", in Charles J. Petrie Jr., ed., Enterprise Integration Modeling: Proceedings of the First International Conference, MIT Press, Cambridge, MA, 1992.
- [HUHN94] Huhns, Michael N. , Munindar P. Singh, Tomasz Ksiezzyk, and Nigel Jacobs, ``Global Information Management via Local Autonomous Agents,’’ *Proceedings of 13th International Workshop on Distributed Artificial Intelligence*, Seattle, WA, August 1994.
- [KOST94] Koster, Martijn, "ALIWEB - Archie-Like Indexing in the WEB", First International Conference on World-Wide, May, 1994.
- [SING94] Singh, Mona, Munindar Singh, and Darrell Woelk. "Knowledge-Based Natural Language Interfaces for Information Access", MCC Technical Report *in preparation*.
- [TOML92] Tomlinson, C., G. Lavender, G. Meredith, D. Woelk, and P. Cannata. "The Carnot Extensible Services Switch (ESS) - Support for Service Execution," in Charles J. Petrie Jr., ed., Enterprise Integration Modeling: Proceedings of the First International Conference, MIT Press, Cambridge, MA, 1992.
- [WAIS] <http://info.cern.ch/hypertext/Products/WAIS/Overview.html>
- [WOEL92] Woelk, D., W. Shen, M. Huhns, and P. Cannata, "Model Driven Enterprise Information Management in Carnot", in Charles J. Petrie Jr., ed., Enterprise Integration Modeling: Proceedings of the First International Conference, MIT Press, Cambridge, MA, 1992.
- [WOEL93] Woelk, D., P. Cannata, M. Huhns, W. Shen, and C. Tomlinson. "Using Carnot for Enterprise Information Integration". Second International Conference on Parallel and Distributed Information Systems. January 1993. pp. 133-136.

Author Biographies

Darrell Woelk is Director of the Carnot Project at MCC. He graduated from the University of Kansas with a B.S. in Engineering Physics and from Kansas State University with an MS in Computer Science. He worked for NCR for 21 years in various development organizations, including Advanced Development at NCR in the areas of database research speech recognition, and optical storage. He was assigned to MCC from NCR in 1985. At MCC, he was one of the developers of the ORION distributed object-oriented database system and the Carnot enterprise information integration system. He is the author of numerous papers on query processing, multimedia information management, authorization in object-oriented database systems, heterogeneous database management, and advanced transaction management. He holds a patent for a unique database machine architecture.

Christine Tomlinson holds a B.A. in applied mathematics and psychology from Rice University. She was employed by Burroughs Corporation first in systems support for large multiprocessor

systems and then in research and development in the Federal and Special Systems Group focusing on local area network technologies and distributed and secure operating systems projects. From this assignment she became the principal architect for a distributed office systems product development. After thirteen years with Burroughs, Ms. Tomlinson joined the Parallel Processing Program at MCC in 1986 where she developed a series of language systems aimed at parallel processing problems in the domain of symbolic data. She is the principal developer of the Extensible Services Switch technology and has played a key role in the development of the Carnot architecture. She holds four patents in areas related to microprocessor design and has published a variety of papers in areas related to object-oriented technology.

woelk@mcc.com