Research in Multi-Organizational Processes and Semantic Information Brokering at the LSDIS Lab

Amit Sheth, John Miller, Krys Kochut, Budak Arpinar

Large Scale Distributed Information Systems (LSDIS) Lab
Department of Computer Science - The University of Georgia
415 Graduate Studies Research Center, Athens, GA 30602-7404 USA
http://lsdis.cs.uga.edu
Phone: +1-706-542-2310; Fax: +1-706-542-2966
{amit, jam, kochut, budak }@cs.uga.edu

1 Introduction

The Large Scale Distributed Information Systems Lab (LSDIS) at the University of Georgia was established in the Fall of 1994 to perform research and technology development on two complementary themes in distributed information systems: **Enabling Infocosm** and **Processes for Networked Organizations**.

Enabling Infocosm Theme. The merging of computers and communications with significant advances in networking infrastructure has made millions of information sources with wide varieties of information accessible. This capability gives us a vision of an information rich society "infocosm", where we expect to have *any information any where we want in (m)any form(s) for effective decision making and knowledge-centric activities, improved productivity, and fun*. Emphasis of our research is on semantics (meaning and use) of information and semantic interoperability supported by an information brokering architecture. This is closely tied to the recent interest in Semantic Web, which is presented as "the Web of data (and connections) with meaning in the sense that a computer program can learn enough about what the data means to process it (Tim Berners-Lee, Weaving the Web, 1999). Our emphasis is on *organizing and utilizing Web-based as well as enterprise content semantically,* rather than though syntactic and structural methods.

Research in enabling Infocosm through semantic information brokering is anchored by the InfoQuilt, project, with InfoHarness/VisualHarness, ADEPT, and Video Anywhere as associated projects (Section 2).

Processes for Networked Organizations. The interoperability of information systems, applications and users within and across enterprises has important implications on the competitiveness of organizations in the global economy. Three important components of a comprehensive approach to enterprise integration are: integration and interoperability of applications and information systems (as in Enterprise Application Integration), coordination of activities through automated process and workflow management, and collaboration between humans as well as effective interactions between humans, applications and information systems. Primary project in this theme has been METEOR, with CaTCH as an associated project, and research as well as applications in Bioinformatics (Section 3).

An important aspect of LSDIS's activities is that research, technology development and prototyping efforts have occurred in close collaboration with industry and government partners. We strive to achieve impact of our research not only through publications and invited talks, but also through industry trials, real-world applications, technology transfers and commercialization. Our significant success in this respect is summarized in Section 4.

2 Enabling Infocosm: InfoQuilt and related projects

InfoQuilt's objective is to enable Infocosm, where we can support effective decision making and improve productivity through a variety of semantic techniques to organize and utilize all forms of data and information, on the Web and within an enterprise. The InfoQuilt system uses a multi-agent information brokering architecture to support the following capabilities:

- Access, analyze and interoperate with heterogeneous, static or dynamic (pull and push) content (pages, documents, sites, repositories, databases) using both wrapper and information extraction technologies
- Semi-Automatically or Automatically create semantic metadata (domain-specific or

contextually relevant metadata), as well as the syntactic metadata

- Use and support of multiple (possibly preexisting) ontologies and domain modeling with complex inter-ontology relationships, domain rules and functional dependencies
- User defined functions (esp. for fuzzy/approximate matching), simulation; also used in post processing result analysis
- Information request processing utilizing domain semantics and resource characteristics (local completeness, data characteristics, binding patterns)

At the conceptual level, we have been interested in representation of context and semantic proximity [Kashyap and Sheth 1996], use of semantic metadata and semantic information brokering architecture [Kashyap and Sheth 2000], and the computation of information loss in multi-ontology query processing [Mena et al 2000]. Our recent interest is in supporting knowledge discovery from heterogeneous, autonomous sources of information, and supporting learning through what-if analysis of empirically defined relationships between data. To this end, we have introduced the computing paradigm of ISCape (Information Scape) which allows users to query and analyze the data available from a diverse autonomous sources, gain better understanding of the domains and their interactions as well as discover and study relationships. A sample of IScape used as an Information Request is shown in Figure 1.

Use of IScape for knowledge discovery can be exemplified using the following investigation [Thacker et al 2001]:

"Do Nuclear Tests result in Earthquakes?" Here a researcher has a hypothesis that this is the case, and wants to prove or disprove it by studying independently collected data made available at various Web-accessible information sources. The knowledge discovery process involves use of ontologies related to Nuclear Tests and Earthquakes, and complex relationships such as "causes" with spatial and temporal parameters.

Several projects are precursors to the InfoQuilt project.

InfoHarness/VisualHarness

The InfoHarness system (1994-1997) [Shah and Sheth 1999] emphasized use of metadata in search and retrieval of heterogeneous information in intranet/Internet environments. The basic InfoHarness system, provided rapid access to huge amounts of heterogeneous information on the Web and corporate intranets, without reformatting, restructuring, or relocating the data. It also supported logical restructuring of the information space, support for multiple third-party indexing engines and many other features. Research in the joint Bellcore-LSDIS project addressed several new research issues, including: keyword as well attribute-based access to document collections, access to multiple autonomous repositories with independent and heterogeneous indices for scaleable search with associated intelligent merging of results, access to the same repository with multiple indices to improve quality of result, and support for remote servers using CORBA.

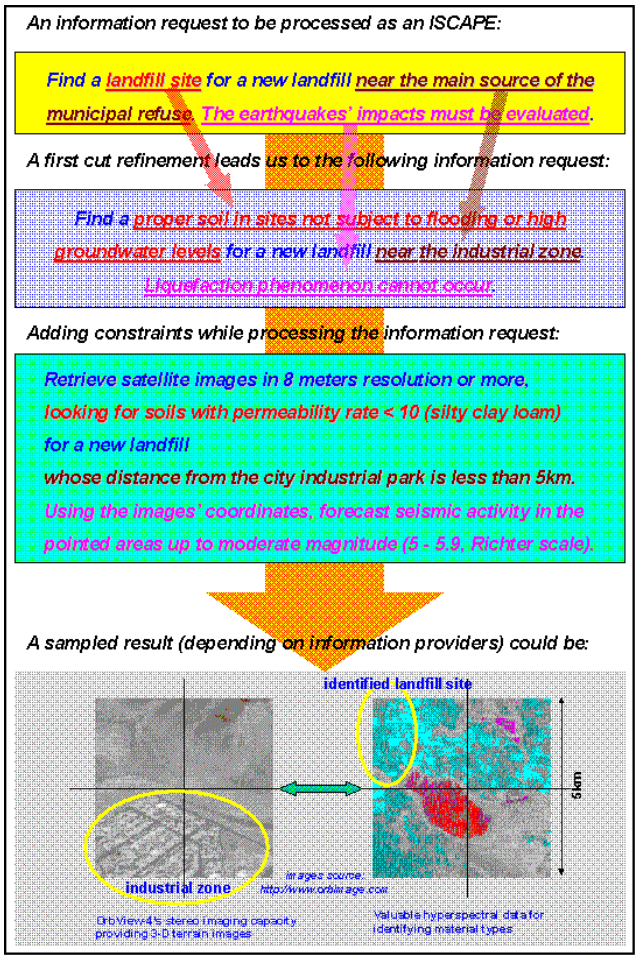

Figure 1: IScape for Information Correlation and Request

VisualHarness system (1997-1998) was the InfoHarness system enhanced to provide customizable and extensible search of federated image repositories, in addition to text and semi-structured data, and structured (relational) databases. Through the comprehensive use of

metadata of various types, it supported keyword, attribute-based and content-based search over images. Also, different Visual Information Retrieval (VIR) engines and third party indexing technologies could be hooked into the system. A key idea was to form a combination of these different access strategies to achieve better quality results. One interesting technique developed was the black box approach for extracting the required information from the VIR engine in the form we could access (distance metrics), without knowing the internals of the VIR engine. We have shown the validity of this black box approach by comparing the results obtained with this approach against the results obtained from a VIR engine [Sheth et al 1999].

VideoAnywhere
This industry-funded project researched the issues related to management of video in embedded (e.g., cable set top) system running a Java Virtual Machine. It targeted interactive TV and video application market-place. It involved research in understanding metadata and consumer profile, caching, crawling (using agent technology), etc. We also prototyped a Web-based search engine for streaming audio and video on the Web, premium TV programming, and local video (such as DVDs).

ADEPT$_{UGA}$
UGA´s work on ADEPT - Alexandria Digital Earth Prototype (1999-2000) involved: (a) specifying metadata for geospatial information; (b) prototyping IScape to support Digital Earth metaphor, incorporating complex and inter-ontological relationships like “affects” and user defined functions including simulations, and (c) development of a graphical knowledge management toolkit for easy creation and deployment of IScapes, Ontologies and Relationships.

Looking ahead, our research in multi-agent semantic information brokering is targeted at semantic-level solutions to serve a variety of information stakeholders (participants), support learning and decision-making in a more comprehensive manner, ease human – information system interaction through semantics (such as better understanding of user’s context and information need), and develop next generation of content management regardless of source, format, media, representation and modeling of content.

3. Enterprise and Multi-organizational Processes

Workflow management is the automated coordination, control and communication of work as is required to satisfy organizational processes. A Workflow Management System (WFMS) is a set of tools providing support for the necessary services of workflow process creation, workflow enactment, and administration and monitoring of workflow processes. The METEOR system addresses the challenges of demanding multi-organization processes and focuses on coordination of user and automated tasks in real-world multi-enterprise heterogeneous computing environments.

METEOR [Managing End-To-End OpeRations]
The METEOR system consists of a suite of four components (Figure 1):

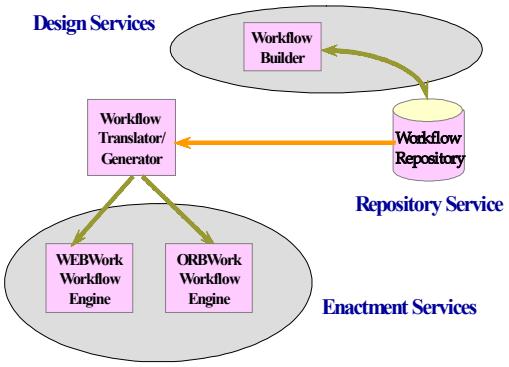

Figure 1. METEOR System Architecture

- Builder for comprehensive, fully graphical process modeling
- Repository for workflow component browsing, reuse and sharing
- A choice of interoperating Enactment Services:
  - WebWork a fully distributed, Web based service [Miller et al 1998]
  - ORBWork for fully distributed, CORBA based service [Kochut et al 1999] with recent support for Java/RMI distributed infrastructure instead of CORBA
- Manager for monitoring/administration

A unique aspect of the METEOR system is automatic code generation for a heterogeneous, distributed environment from graphical design/ specification. This can lead to 60 to 90% saving in enterprise process application development and deployment, with corresponding savings in costs and time.

METEOR has continually evolved as we studied the requirements of healthcare, defense telecommunications and e-commerce applications with our industry partners. Recent focus has been on supporting the following capabilities:

High Security: Naval Research Lab (NRL) and LSDIS Lab worked together to provide security within and across security domains (authentication, authorization, access control, encryption and non-repudiation). Multilevel Security (MLS) features are now part of the METEOR/NRL Workflow Builder and the ORBWork enactment service.
High Reliability and Survivability: METEOR's ingredients for achieving high reliability and survivability for workflows are exception handling, recovery, and adaptability. Recent research in exception handling has involved support for cross-organizational processes using case-based reasoning, knowledge sharing, coordination, and intelligent problem solving.
Scalability: Once a workflow has been developed, it needs to be deployed, possibly over a large geographical region and including multiple organizations. ORBWork's fully distributed architecture and implementation, utilizing CORBA, Web and/or Java technologies, is highly scalable. ORBWork has configuration specifications that make it easy to deploy workflow applications, as well as move or relocate tasks at run time for performance improvement and adaptation.
Comprehensive Modeling: We have investigated the relationships of workflow modeling with database and simulation modeling [Miller, Sheth, Kochut 1999].
Rapid Design, Development & Deployment: We have developed a comprehensive repository (based on XML and object-relational database technology) to store metadata about workflow designs, organizations, informational resources (e.g., application databases) and computational resources (CORBA servers, Java Servlets, or EJB components). The repository enhances reusability and facilitates rapid incorporation of resources into workflows, as well as supports adaptability.
Adaptability: Today's military/commercial environments may rapidly undergo major changes. Workflows must adapt to these changes. The built-in adaptability of ORBWork combined with an easy-to-use graphical designer, and a comprehensive repository make rapid adaptation possible. In addition, monitoring tools as well as advanced exception handling mechanisms are used to help indicating when adaptation may be necessary.
Inter-organizational Workflows: The problem of multi-workflow enactment has many facets, ranging from technological issues about how to integrate different workflow management systems of different vendors on different platforms to the purely conceptual issues of specifying how the interaction should occur, i.e., semantic interoperability. The enactment of different workflows, which can be autonomously and separately designed, may cause problems such as deadlocks, starvation, live-locks, or failure to terminate in the desired final state. The available interoperability specifications, such as SWAP and JFlow help in providing multi-workflows; but they are still far from meeting interoperability needs of workflow systems.

Our current research directions involve development of a multi-agent framework to support advanced capabilities for inter-organizational workflows and Web services. Some current issues we are investigating include Quality of Service specification, and survivability techniques for federated Web services besides use of workflow techniques for Web service integration and management.

CaTCH

While focus of METEOR has been on coordination, collaboration is also a very important part of enterprise processes, and was investigated in the CaTCH (Collaborative Teleconsulting for Healthcare) project. The CaTCH project involved integration for several rapidly progressing technologies to support high bandwidth collaboration through access to diverse variety of information over a variety of communications alternatives. CaTCH I supported real-time collaboration by integrating multimedia patient data on Intranet, medical reference data on Internet, LAN/POTS/ISDN-based Video+Data Conferencing, and WWW/Java programming to set up remote environment and context sensitive collaboration. CaTCH II supported asynchronous (store-and-forward) collaboration using a Web-object model to integrate a variety of information such as streaming video of medical information, video or audio mail, patient data, etc. and making it available on demand through Web.

Bio-informatics Research and Applications

Bio-informatics research constitutes an important area in which LSDIS contribute extensively, especially in management of genomics data and automating complex workflows in genome labs. In a joint project with UGA Genetics Department, LSDIS is working to develop IntelliGEN, a comprehensive system for genomic data and process management. An important goal of this project is to discover

interactions and roles of proteins in an organism (i.e. *Neurospora crassa).*

Manual handling of plate tracking, data collection, and computation of the protein-protein interaction map processing of the planned 500 bait library plates is extremely labor intensive and error-prone. In addition, most of the genomics and bioinformatics tasks involved in the proposed project are very long and complex. To address this problem, we are building upon our extensive work on an extensible data and process support management system for handling the experiments in protein-protein interaction mapping.

We have developed an automated workflow system called GENEFLOW [Hall et al 2001] as well as the Fungal Genome Database (FGDB) in order to build a laboratory information system for managing distributed high throughput sequencing. In addition, we have created graphical tools to visualize the mapping and sequencing data. These existing graphical database tools support XML messaging to exchange genomic information with other databases and applications. While earlier workflow systems have been used to automate lab experiments, we believe that current advances in adaptive workflow technologies can improve dramatically the quality of experiments by optimizing laboratory workflows.

In the near term, the core objective of the proposed system is running protein-protein interaction workflows. However, we plan to use the system in other types of genomic workflows in the future.

As a conclusion, we believe that most of the attention in Information Systems has gone to data management and interoperability, and this attention will increasingly shift to information and knowledge with support for semantics on the one hand, and processes on the other. The first deals with what is the service and product in E-commerce applications for example, and the second deals with how to effectively support or render it. In our vision partly outlined in [Sheth 1999] and [Sheth, Aalst, and Arpinar 1999], we propose that (a) semantics is the key to effective organization and utilization of Web-based and Enterprise-based information regardless of syntax, format/representation, media, and source, (b) for the processes in future networked economy, in which we will see process as an organic part of doing a business—that is, while processes will be chief differentiating and the competitive force in doing business in the networked economy, they will be deeply integrated with the way of doing business, and that they will be critical components of almost all types of systems supporting enterprise-level and business critical activities.

4. Impact

LSDIS's research has achieved significant impact through several avenues, including training of students, commercialization, and industry collaborations.

Technology Transfer and Commercialization:

- InfoHarness research lead to its commercialization as Adapt/X Harness from Bellcore (now Telcordia Technology) in 1995.
- METEOR research lead to technology licensing to Infocosm, Inc., which then released a commercial version called METEOR EAppS in 1998, and was licensed to enterprises.
- VideoAnywhere technology resulted into Venture Capital funded spin-off Taalee, Inc. in 1999 (recently acquired by Voquette, Inc), which has patented methods related to Semantic Search, Personalization, Directory and Interactive Marketing, commercialized the Semantic Engine™ technology, and licensed corresponding commercial products and services [see *www.taalee.com* and *www.voquette.com*].

Significant Applications and Trials (a partial list):

- Medical College of Georgia (MCG) and LSDIS developed two applications: (a) Pediatric Echocardiograph Consultation application for use by rural physicians to get consultation at MCG using the CaTCH system developed at LSDIS (a field trials proved that this Web based system yielded the same results as those achieved using traditional approach of viewing tapes on a TV, while replacing overnight tape delivery by instant and cost-effective Web based delivery of media and patient information), (b) CareWeb, an advanced Web-based system developed in the Georgia's Family Connections program to support distance health education and collaboration among patients and healthcare workers.
- Infocosm Inc., MCG and Advanced Technology Institute used LSDIS's METEOR system to develop workflow applications to support neonatal clinical pathways.
- LSDIS developed CAPA (Course Approval Process Automation) using some of METEOR technology and LSDIS's expertise in workflow

™ Semantic Engine is a trademark of Taalee, Inc.

management. The system has been operational since 1998. It supports a 30+ step workflow spread over entire UGA campus with user-base of 130+ departments and 3500+ faculty and staff members, and has already been used to support approval of over 6000 courses.

- Connecticut Healthcare Research and Education Foundation (CHREF) and LSDIS developed State-wide Child Immunization Tracking application using the METEOR WebWork system
- METEOR designer has been made available free of charge to instructors and used in conducting advanced graduate courses in this area at several institutions worldwide.

Acknowledgements

Several long term visitors to the LSDIS lab have participated and enriched our research. These include, Professors Tarcisio Lima (Department of Computer Science, Federal University of Juiz de Fora, Brazil), Wil van der Aalst (Eindhoven University of Technology), Eduardo Mena (Universidad de Zaragoza), and Mizuho Iwaihara (Kyoto University). Collaborations with several institutions have played important role in developing research objectives through understanding of real-world requirements and evaluating our prototypes. Key collaborating institutions include CHREF, MCG (Dr. Warren Karp), Advanced Technology Institute (Jack Corley), NRL (Dr. Myong Kang), Telcordia Technology (Dr. Marek Rusinkiewicz), and Medical University of South Carolina. LSDIS research has been funded by many sources, including, NIST (HIIT and HITECC programs), Community Management Staff (Massive Digital Data Systems Initiative), NSF (Digital Library II initiative), NRL, LGERCA, the Boeing Corporation, MCC, and Hewlett-Packard. Significant equipment and software donors include Informix, Microsoft, Iona, POET and Virage.

References

D. Hall, J. Miller, J. Arnold, K. Kochut, A. Sheth and M. Weise, "Using Workflow to Build an Information Management System for a Geographically Distributed Genome Sequencing Initiative," *Genomics of Plants and Fungi,* R.A. Prade and H.J. Bohner, Editors (2001) Marcel Dekker, Inc. New York, NY.

V. Kashyap and A. Sheth, "Schematic and Semantic Similarities between Database Objects: A Context-based Approach," Very Large Data Bases (VLDB) Journal, 5(4), October 1996, pp 276-304.

V. Kashyap and A. Sheth, Information Brokering Across Heterogeneous Digital Data, Kluwer Academic Publishers, August 2000, 248 pages.

K. Kochut, A. Sheth, and J. Miller, "Optimizing Workflow: Using a CORBA-based, Fully Distributed Process to Create Scalable, Dynamic Systems," Component Strategies Journal, March 1999.

E. Mena, V. Kashyap, A. Illarramendi, A. Sheth, "Imprecise Answers in Distributed Environments: Estimation of Information Loss for Multi--Ontology based Query Processing"
the International Journal of Cooperative Information Systems (IJCIS), 9 (4) (2000), pp 403-425.

J. Miller, D. Palaniswami, A. Sheth, K. Kochut and H. Singh, "WebWork: METEOR2's Web-Based Workflow Management System," Journal of Intelligent Information Systems: Integrating Artificial Intelligence and Database Technologies (JIIS), 10 (2) (March/April 1998) pp. 185-215.

J. Miller, A. Sheth and K. Kochut, "Perspectives in Modeling: Simulation, Database and Workflow," *Conceptual Modeling: Current Issues and Future Directions*, LNCS-1565, P. Chen, J. Akoka, H. Kangassalo, B. Thalheim, Eds, (April 1999) pp 154-167. Springer Verlag.

A. Sheth, K. Shah, K. Parsuraman and S. Mudumbai, "Searching Distributed and Heterogeneous Digital Media," 8th IFIP 2.6 Working Conference on Database Semantics (DS-8), Rotorua, New Zealand, January, 1999.

K. Shah and A. Sheth, "InfoHarness: An Information Integration Platform for Managing Distributed, Heterogeneous Information," IEEE Internet Computing, November-December 1999, pp 18-28.

A. Sheth, Changing Focus in Interoperability in Information Systems: From System, Syntax, Structure to Semantics, Interoperating Geographic Information Systems, in Interoperating Geographic Information System, M. Goodchild, M. Egenhofer, R. Fegeas, and C. Kottman, Eds, Kluwer Academic Publishers, 1999.

A. Sheth, W. M. P. van der Aalst, and I. B. Arpinar, Processes Driving the Networked Economy: Process Portals, Process Vortexes, and Dynamically Trading Processes, IEEE Concurrency Journal, July-September 1999.

S. Thacker, A. Sheth and S. Patel, "Complex Relationships for the Semantic Web," Creating the Semantic Web, D. Fensel, J. Hendler, H. Liebermann, and W. Wahlster (eds.) MIT Press, 2001 (in print).

Additional project details and references can be found at: *http://lsdis.cs.uga.edu*.

A list of keynotes and invited talks with associated presentations that further outline our vision is at *http://lsids.cs.uga.edu/~amit*.