Large Scale Distributed Information Systems (LSDIS) Lab, University of Georgia
Research in Semantic Bioinformatics, Semantic Analytics and Semantic Web Processes. http://lsdis.cs.uga.edu

The LSDIS Lab advances the field of distributed information systems by researching semantic techniques for exploiting heterogeneous multimedia information and improving processes, encompassing the central promise of the Semantic Web initiative. It is pursuing cutting edge research in ontology development for demanding scientific domains, semantic heterogeneity and integration, complex relationships discovery and semantic analytics, and Semantic Web services and processes.   Past work of the LSDIS lab can be characterized by keywords: semantic interoperability, syntactic and semantic metadata for text and digital media, metadata based integration of Web content, ontology-driven information systems, multi-ontology query processing, transactional workflows and workflow management.   Significant past projects include InfoHarness/VisualHaness, InfoQuilt, OBSERVER, and METEOR.

The LSDIS Lab was established in 1994 with the guidance and direction provided by Dr. Amit P. Sheth with the help of Dr. John A. Miller and Dr. Krzysztof J. Kochut. In 1998, this faculty group was further strengthened by the addition of Dr. I. Budak Arpinar.   Currently, the lab consists of fifteen research assistants, majority of them pursuing PhD., and one research staff.

LSDIS Lab has delivered on innovation in research with impact on the real world. The LSDIS library consists of a large number of highly cited publications in top journals and conferences. Strong symbiotic relationships have been developed with different industry partners, leading to commercial licenses and technology transfer, patents and contribution to standards activity. The "Video Anywhere" project was licensed to Taalee, Inc. (now Semagix Inc. which offers comprehensive Semantic Web technology based Freedom product for building search, integration and analytics applications), and the METEOR project was licensed to Infocosm, Inc, resulting the METEOR-EAppS product. WSDL-S, a proposal for adding semantics to WSDL, has been proposed to the W3C WSDL committee and our collaborators at the IBM T.J. Watson Research center.   The distinguishing feature of our research is that it is often multi-disciplinary, whether exploiting multiple subfields of computer science or working with researchers and professionals in other fields such as bioinformatics, geographical information systems, national and homeland security and healthcare informatics.

The ongoing projects as LSDIS Lab are listed in alphabetical order:-

Glycomics (Bioinformatics for Glycan Expression)
http://lsdis.cs.uga.edu/Projects/Glycomics/

This project is one of the four component of a large National Institute of Health funded Center on ""*Integrated Technology Resource for Biomedical Glycomics,"* and involves collaboration with the LSDIS lab and the Complex Carbohydrate Research Center at the University of Georgia (Prof. William York). Its objective is to develop a suite of databases along with computational tools that facilitate efficient acquisition, description, analysis, sharing and dissemination of the data contained therein. This represents a major challenge, as the potential of this data to explain important biological phenomena will only be fully realized if it is examined in the context of the vast amounts of other data that are becoming available. Therefore, a major emphasis is placed on data structures and tools that have a high degree of interoperability with the computational

infrastructure now being developed for the storage and analysis of genomics and proteomics data. The specific aims of this research are as follows.

- Develop and implement efficient workflow tools for tracking physical samples and for automating data collection, data verification, compression, and storage. These will include tools for automatic identification of glycan structures and/or glycan structural families from mass spectral data.
- Build an integrated database termed GlycomBin that describes the populations of specific glycan structures and structural families of glycopeptides and glycolipids in different cell lines.
- Develop tools that facilitate interoperability of the databases with existing proteomics tools that can be used, for example, to identify and quantitate the expression level of each glycopeptide's parent protein and the expression levels of the proteins involved in glycan biosynthesis. Support open standards-based access of GlycomBin and its interoperability with external databases. This will include Web Services enabled access to data and computational resources.
- Develop tools that facilitate the description, classification, and clustering of glycopeptides and glycolipids, including ontology based semantic descriptions of glycan structure, biosynthesis, and biological context. This requires the development of a set of interdependent ontologies, called GlycO (for "Glycomics Ontology"). GlycO is populated with extensive knowledge that embodies semantically rich descriptions of the domains of carbohydrate structure, glycan binding relationships, glycan biosynthetic pathways, and the developmental biology of stem cells. Classes of objects and their relationships in GlycO model information we store about the differential expression of glycan structures on the surface of developing stem cells. We are developing methods to automate the population of these ontologies from multiple, heterogenous (semi-structured and structured) knowledge sources. For example, the structure ontology is populated with specific glycan structures and the building blocks (glycosyl residues) from which these molecules are assembled. Provenance, i.e., the sources of this information and the reliability of those sources, is also incorporated into the ontology description and the knowledge base.
- Develop a new knowledge representation paradigm that allows semantic representation of partial or incomplete knowledge as well as fuzzy set membership. In scientific domains it is not only important to deductively derive implicit knowledge from the explicit facts in the ontology, but also to test the likelihood or the degree of truth of a hypothesis based on the current state of knowledge, which is mostly incomplete.
- Develop tools for semantic data analysis and discovery, including tools for finding correlations between glycosylation patterns and patterns of gene expression within a cell line or between different cell lines. These will include a blended ontology-supported browsing and querying interface.
- Develop methods for automatic semantic annotation of scientific data resulting from experimentation. Example of data to be annotated include mass spectrometry. Novel aspect of this work includes development of a process ontology (termed GlycoPro) for formal specification of experimental process and to annotate scientific data with both domain and process ontology. This approach significantly extends typical provenance information of biological data, leading to more opportunities for mining and discovery.

This approach will provide a highly flexible environment for the development of distributed and semantic bioinformatics approaches for analysis of glycosylation patterns and their biological relevance The project web site provides the status of openly available results including a version

of Glyco and GlycoPro, a complex carbohydrate data interchange format, a tool for graphical display and provenance of ontologies, as well as recent presentations and publications.

This research is primarily funded by "Bioinformatics of Glycan Expression," with is one of the four components of the NIH sponsored center "Integrated Technology Resource for Biomedical Glycomics," (center PI: Mike Pierce, CCRC; funding: appox. $6 million, National Institute of Health, July 1, 2003 – June 30, 2008).

METEOR-S: Semantic Web Services and Processes

Applying Semantics in Annotation, Quality of Service, Discovery, Composition, Execution
http://lsdis.cs.uga.edu/Projects/METEOR-S/

The METEOR project at the LSDIS Lab, University of Georgia, focused on workflow management techniques for multi-paradigm transactional workflows. METEOR-S is the follow on project, which supports Web-based business processes within the context of Service Oriented Architecture (SOA) and the semantic Web technologies and standards. Rather than reinvent from scratch, METEOR-S attempts to build upon existing SOA and Semantic Web standards whenever possible (using extensibility features) where appropriate, or seeks to influence existing standards to support and exploit semantics. WSDL-S [Miller et al., 2004], a proposal for adding semantics to WSDL, has been submitted to the W3C WSDL committee and our collaborators at IBM T.J. Watson Research center.
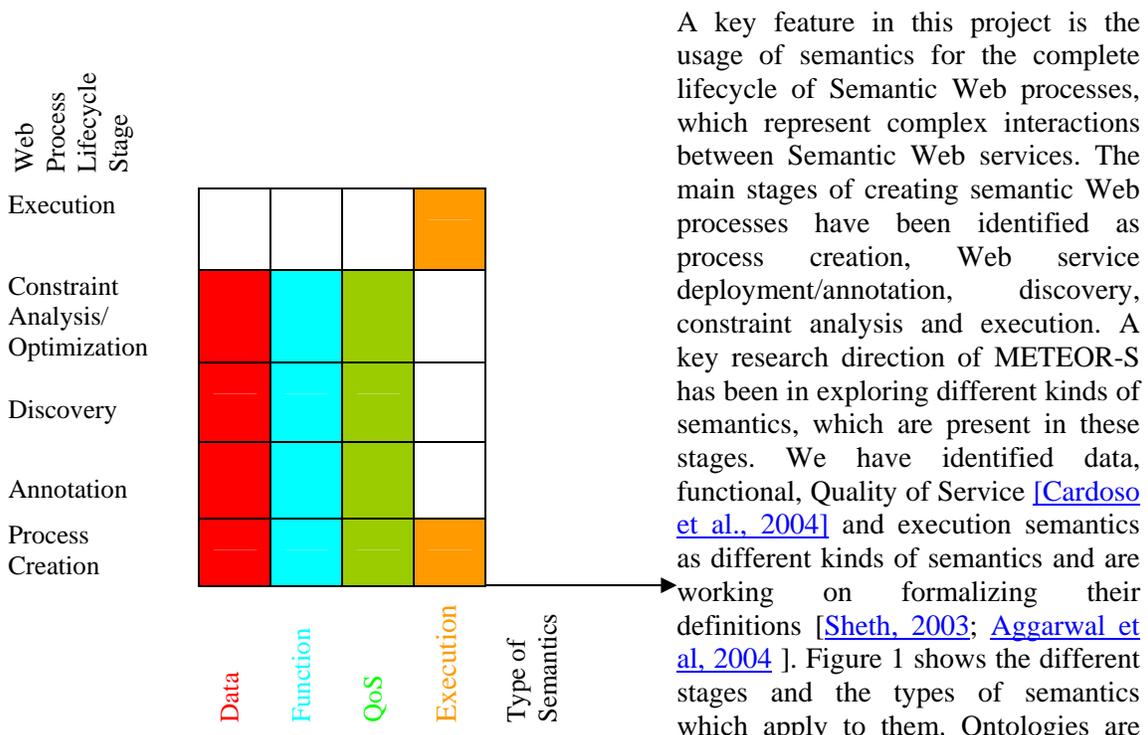

Figure 1: Different Stages of Web process lifecycle and different types of semantics

A key feature in this project is the usage of semantics for the complete lifecycle of Semantic Web processes, which represent complex interactions between Semantic Web services. The main stages of creating semantic Web processes have been identified as process creation, Web service deployment/annotation, discovery, constraint analysis and execution. A key research direction of METEOR-S has been in exploring different kinds of semantics, which are present in these stages. We have identified data, functional, Quality of Service [Cardoso et al., 2004] and execution semantics as different kinds of semantics and are working on formalizing their definitions [Sheth, 2003; Aggarwal et al, 2004]. Figure 1 shows the different stages and the types of semantics which apply to them. Ontologies are the primary mode of expressing and reasoning on the various kinds of semantics.

The key components of METEOR under various stages of development are

1. METEOR-S Web Service Annotation Framework
   This project provides a GUI based tool which enables users to semi-automatically annotate Web services with existing ontologies. It uses a variety of matching algorithms and techniques for predicting the matches. Some of the algorithms /techniques used are
   - NGram for linguistic similarity
   - Exploitation of relationships in wordnet
   - Abbreviation dictionary

   Further details of these algorithms are present in the WWW2004 paper [Patil et al., 2004]. It is currently available for download at http://lsdis.cs.uga.edu/Projects/METEOR-S/MWSAF/

2. METEOR-S Publication and Discovery Interface
   This project provides an eclipse based toolkit for semantic publication and discovery of Web services. Users can annotate their Web services either manually [Rajasekaran et al., 2004] or using the matches suggested by MSWAF. This tool provides a GUI to publish the annotated Web services. It also provides an interface for discovery [Verma et al., 2004a] with the help of semantic templates. This tool is currently in the testing stage and will be released in October, 2004.

3. METEOR-S Dynamic Composition Environment
   This process leverages semantic annotation, publication and discovery of Web services to create an environment for dynamic composition of Web services [Aggarwal et al., 2004]. Users can specify process QoS constraints, domain constraints, inter service dependencies [Verma et al., 2004b] and service templates for the services. The constraint analyzer module produces an optimal set of services to be bound to the process. This tool is under late stages of development and will be released in late October 2004.

REFERENCES

[Aggarwal et al, 2004] Aggarwal, K. Verma, A. Sheth, J. Miller, W. Milnor, Meteor-S Dynamic Composition Environment, The proceedings of 2004 IEEE International Conference on Services Computing, September 2004.

[Cardoso et al., 2004] J. Cardoso, A. Sheth, J. Miller, J. Arnold, and K. Kochut, Quality of Service for Workflows and Web Service Processes, Journal of Web Semantics, Elsevier, 1 (3), 2004, pp. 281-308.

[METEOR-S] METEOR-S: Semantic Web Services and Processes, http://lsdis.cs.uga.edu/Projects/METEOR-S/

[Miller et al., 2004] J. Miller, K. Verma, P. Rajasekaran, A. Sheth, R. Aggarwal, K. Sivashanmugam, WSDL-S: A Proposal to W3C WSDL 2.0 Committee, LSDIS Lab, June 2004.

[Patil et al., 2004] A. Patil, S. Oundhakar, A. Sheth, K. Verma, METEOR-S Web service Annotation Framework, The proceedings of the 13th International World Wide Conference, (2004).

[Rajasekaran et al, 2004] Enhancing Web Services Description and Discovery to Facilitate Orchestration, Proceedings of the First International Workshop on Semantic Web Services and Web Process Composition. (SWSWPC 2004), July 2004, pages 34-47.

[Sheth, 2003] A. Sheth, "Semantic Web Process Lifecycle: Role of Semantics in Annotation, Discovery, Composition and Orchestration," Invited Talk, Workshop on E-Services and the

Semantic Web (co-located with World Wide Web Conference, 2003), Budapest, Hungary, May 20, (2003).

[Sivashanmugam et al., 2003] K. Sivashanmugam, K. Verma, A. Sheth, J. Miller: Adding Semantics to Web Services Standards, Proceedings of 1st International Conference of Web Services, 395-401, (2003).

[Verma et al., 2004a] K. Verma, K. Sivashanmugam, A. Sheth, A. Patil, S. Oundhakar and J. Miller, METEOR–S WSDI: A Scalable Infrastructure of Registries for Semantic Publication and Discovery of Web Services, Journal of Information Technology and Management (in print), (2004).

[Verma et al., 2004b] K. Verma, R. Akkiraju, R. Goodwin, P. Doshi, J. Lee, On Accommodating Inter Service Dependencies in Web Process Flow Composition, AAAI Spring Symposium PP: 37-43 on Semantic Web Services.

SemDis: Discovering Complex Relationships in Semantic Web

http://lsdis.cs.uga.edu/Projects/semdis/

The Semantic Discover (**SemDis**) project is focused on the design, prototyping and evaluation of a system that supports indexing and querying of complex semantic relationships and is driven by notions of information trust and provenance and models of hypotheses and arguments under investigation. Instead of a search engine that returns documents containing terms of interest, we are developing a system that returns actionable information (with the associated sources and supporting evidence) to a user or application. From a real world application perspective, the discovery of semantic relationships is highly relevant in the domain of national security intelligence [Sheth et al., 2005].
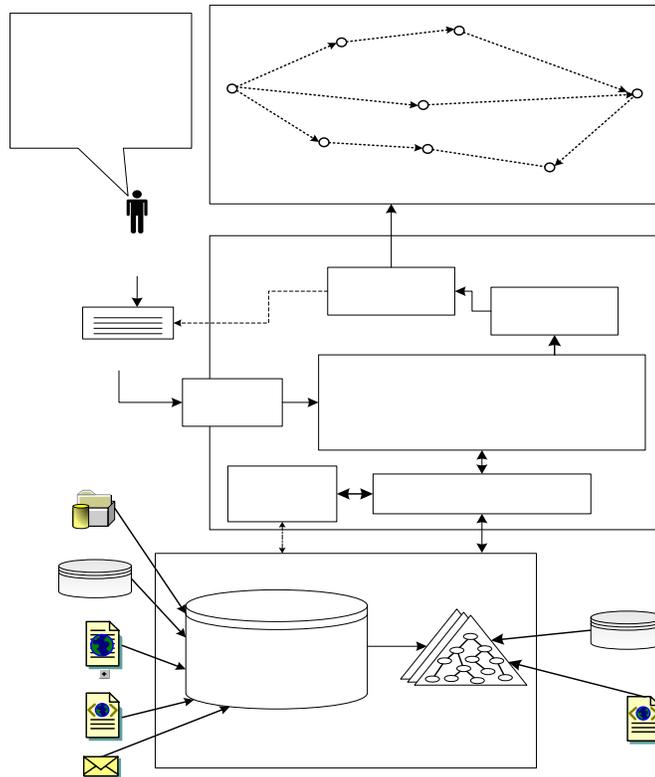

Figure 2: SemDis Architecture

One of the components of SemDis is the ongoing development and maintenance of a test-bed dataset called **Semantic Web Technology Evaluation Ontology (SWETO)**. It is used in the discovery of *Semantic Associations* (described in [Anyanwu and Sheth, 2003]), and also as a public benchmark for future Semantic Web applications and technologies. Of particular interest is not just the schema of the ontology, but also the population (instances, assertions or description base) of the ontology. SWETO is a populated ontology [Aleman-Meza et al., 2003b]. Hence, it is critical for core semantic issues such as semantic disambiguation as well as being necessary for checking the scalability of tools and techniques (e.g. reasoning techniques). SWETO is available to the Semantic Web community researchers and academics under a Creative Commons license.

We have developed algorithms that discover semantic associations over a large semantic meta-base represented in RFD(S) or OWL. Initial prototypes included depth-first search, and random-walks algorithms, as well as novel algorithms based on Tarjan's algorithm. Several key areas of research being conducted to enhance the performance and scalability of semantic association discovery are described below.

Our work under the title **RDF Store** will support research in the design and implementation of RDF (Resource Description Framework) traversal and querying algorithms used in knowledge discovery. It is a specialized, main-memory storage representation for large ontologies represented in RDF (in the order of 10 million triples), and is optimized for performance and efficiency in retrieving information from such ontologies.

Our research in developing and prototyping **Look Ahead Evaluation of Path Queries in Labeled Directed Graph Models** evaluates path queries using Gaussian elimination techniques to compute path summaries for paths found as a result of a query. A pipelined algorithm is then used to iteratively extract the highest K ranked paths from the summary. This approach obviates the need for the entire graph to be in memory, and avoids the explicit manipulation of all the individual paths at a single time. It also supports the notion of Look-Ahead query processing by computing, with minimal additional cost, partial or complete answers for potential future queries in the same neighborhood as the query that is currently being evaluated.

Our research in **Discovering Context and Semantic Associations in Undirected Edge-Weighted Graphs** uses a fast heuristic algorithm to compute the most contextually relevant paths (semantic associations) between given source and destination entities in an undirected RDF graph. A favorable side effect of the algorithm is that it allows the computation of context (as opposed to requiring it be specified a priori), which is then represented as a regular expression. We are also working on a preliminary version of a semantic index that indexes the resulting regular expressions.

We employ **ranking techniques** to provide users with the most interesting and relevant results [Aleman-Meza et al., 2003a]. Criteria used in ranking include association length, rarity, popularity, trust, context, and subsumption. Issues of trust are being addressed by our collaborators at UMBC [http://semdis.umbc.edu/].

REFERENCES:
[Sheth et al., 2005]   A. Sheth, B. Aleman-Meza, I. B. Arpinar, C. Halaschek, C. Ramakrishnan, C. Bertram, Y. Warke,   K. Anyanwu, D. Avant, F. S. Arpinar, and K. Kochut. Semantic Association Identification and Knowledge Discovery for National Security Applications, Journal of Database Management, 16(1), 33-53, Jan-March 2005.
[Halaschek et al., 2004]   C. Halaschek, B. Aleman-Meza, I. B. Arpinar, A. Sheth. Discovering and Ranking Semantic Associations over a Large RDF Metabase, 30th Int. Conf. on Very Large Data Bases, August 30 - September 03, 2004, Toronto, Canada. Demonstration Paper.
[Aleman-Meza et al., 2003a]   B. Aleman-Meza, C. Halaschek, I. B. Arpinar, A. P. Sheth, Context-Aware Semantic Association Ranking, Workshop, Proceedings of SWDB'03, The first

Sheth et al, "LSDIS Lab: Research in Semantic Bioinformatics, Semantic Analytics and Semantic Web Processes," Semantic Web Research Center Report, SIGSEMIS Bulletin, Vol. 1, No. 3, November 2004.

International Workshop on Semantic Web and Databases, Co-located with VLDB 2003, Berlin, Germany, September 7-8, 2003, pp. 33-50.

[Aleman-Meza et al., 2003b]   B. Aleman-Meza, C. Halaschek, A. P. Sheth, I. B. Arpinar, G. Sannapareddy. SWETO: Large-Scale Semantic Web Test-bed, Proc. of the 16th Intl. Conf. on Software Engineering & Knowledge Engineering (SEKE2004): Workshop on Ontology in Action, Banff, Canada, June 21-24, 2004, pp. 490-493.

[Anyanwu and Sheth, 2003]   K. Anyanwu and A. P. Sheth. ρ-Queries: Enabling Querying for Semantic Associations on the Semantic Web. The Twelfth International World Wide Web Conference, Budapest, Hungary. May 2003.

[Anyanwu and Sheth, 2002]   K. Anyanwu and A. Sheth. The ρ Operator: Discovering and Ranking Associations on the Semantic Web, SIGMOD Record, Vol. 31, No. 4, December 2002, pp. 42-47.