A Methodology for Semantic Integration of Metadata in Bioinformatics Data Sources

Lei Li, Roop G. Singh, Guangzhi Zheng,
Art Vandenberg, Vijay Vaishnavi
Georgia State University
Atlanta, GA 30302
lli@gsu.edu

Sham Navathe
Georgia Institute of Technology
Atlanta, Georgia 30332
sham@cc.gatech.edu

ABSTRACT
Semantic heterogeneity is becoming increasingly prominent in bioinformatics domains that deal with constantly expanding, dynamic, often very large, datasets from various distributed sources. Metadata is the key component for effective information integration. Traditional approaches for reconciling semantic heterogeneity use standards or mediation-based methods. These approaches have had limited success in addressing the general semantic heterogeneity problem and by themselves are not likely to succeed in bioinformatics domains where one faces the additional complexity of keeping pace with the speed at which data and semantic heterogeneity is being generated. This paper presents a methodology for reconciliation of semantic heterogeneity of metadata in bioinformatics data sources. The approach is based on the proposition that by globally monitoring, clustering, and visualizing bioinformatics metadata across disparately created data sources, patterns of practice can be identified. This can facilitate semantic reconciliation of metadata in current data and mitigate semantic heterogeneity in future data by promoting sharing and reuse of existing metadata. To instantiate the methodology, a research architecture, MicroSEEDS, is presented and its implementation and envisioned uses are discussed.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: *Clustering.* This is H.2.5 [**Heterogeneous Databases**]: Data translation.

General Terms
Design

Keywords
Information integration, semantic heterogeneity, metadata, clustering, bioinformatics

1. INTRODUCTION

43rd ACM Southeast Conference, March 18-20, 2005, Kennesaw, GA, USA.

Information sources expand rapidly due to the rapid advances in computer and communication technology [5]. This is even more prominent in the bioinformatics area that now often deals with multi-terabyte constantly expanding datasets from various resources [7] [15] [16].

These bioinformatics datasets are typically from autonomous, sources many of which come from standard formats such as relational model; but several still use flat files, legacy data models or special formats such as ASN.1 which is common for many bioinformatics databases [15]. Most sources are made available on the web through an interface and a login process, but they are free to change schemas, modify content, and post dynamic data reflecting new experiments and discoveries [6]. The volume, diversity, and dynamic characteristic of bioinformatics data sources make the need for effective information integration increasingly evident.

Information integration of bioinformatics data is becoming a very vital problem for the following reasons. For information such as DNA sequences, microarray experimental data or mutational data, scientists are interested "similar" information between different organisms (such as mouse vs. human), or for different parts of the body (e.g., skin cells vs. internal tissue cells), or under different experimental conditions. Another need for data integration across sources and for establishing cross-relationships among data comes from the need to connect sequence data to patient or disease data, to functional information, to biochemical pathway information or to determine interactions between genes and proteins [15].

A key step in information integration is to understand the metadata of data sources [5]. Metadata describes the schema or structure of data. As "data of data" [18], metadata and metadata registries have been widely proposed as a solution to enable greater information sharing [17]. Metadata also acts as critical component for discovering relevance of certain data resources [19].

Metadata plays an even important role in reconciling bioinformatics data sources due to their embedded greater complexity over general data sources [7]. The bioinformatics metadata is often not explicitly or well represented and expands and changes as data sources grow and change. Metadata has been argued as one of the key factors causing heterogeneity of bioinformatics data sources [7]. Let's consider a simple

example. A DNA microarray[1] researcher needs to verify her experiment's results against other microarray experiments which are critical for her future investigation and validation of results [22]. She visits her favorite public repository, ArrayExpress (http://www.ebi.ac.uk/arrayexpress), that stores microarray experimental data but she doesn't find a good match. So, she visits another site, Stanford Microarray Database (SMD) (http://genome-www5.stanford.edu). However, the SMD site uses a different way to describe microarray experiment data, e.g., the type of experiment is referred to as "category" instead of "experiment type" as used in ArrayExpress. She finds more and more variance between ArrayExpress and SMD sites as she explores further. She may spend considerable time in understanding the SMD metadata (the way that SMD site express its microarray data) before continuing her search. This can be a long and tedious process as the researcher examines more microarray sites. Even after exhaustively examining those sites, she may still have questions:

"Did I get a complete set of information about related experiments? Did I fully understand how the new sites describe their data? How do the experimental conditions of my experiment match with the data provided at site XYZ? Has any new information become available after my search? Can an automated process help me discover other microarray data information?"

As illustrated by the above example, understanding metadata is an important "vehicle" to access *data* in a data repository and toward successful data integration. Building on research in the domain of Lightweight Directory Access Protocol (LDAP) metadata [10] [11] [12] [25], this paper outlines a methodology for reconciling and mitigating the semantic heterogeneity of bioinformatics metadata, in particular microarray experimental metadata.

2. RELATED RESEARCH

Similar to the "*accidental*" and "*essential*" complexity of software systems articulated by Brooks [2], heterogeneity of metadata is either of an "*accidental*" or "*essential*" nature. Accidental heterogeneity arises from the use of different formats and representation systems (e.g., XML, flat file, relational database, or other format) and can be solved through translation systems for format conversion. Essential heterogeneity, also called semantic heterogeneity, arises from using varied vocabulary (metadata) to describe similar concepts or data relationships or from using the same metadata to describe different concepts or data relationships [24]. Basic semantic problems may remain even after the accidental heterogeneity is resolved. Semantic heterogeneity is the focus of the research reported in this paper.

A traditional approach to solve the semantic heterogeneity problem is the mediator/wrapper-based strategy [1] [3] [21] [23]. The wrapper stands on top of different data resources and converts the format of a data source based on a unified data model. The mediator runs on top of wrappers and strives to reconcile the semantic conflicts among wrappers. This approach has not been widely successful partly because it is solving the problem reactively after it occurs which is more difficult. Also, as Hernandez and Kambhampati [6] point out, as data sources keep growing and data source formats keep changing, the mediator and wrappers have to be constantly altered correspondingly to provide accurate, up-to-date information.

Centrally controlled standards-based approaches can also in theory overcome the heterogeneity issue. If everybody uses the same terms and follows the same format, the accidental and semantic heterogeneity can be greatly reduced or even eliminated. For example, for the microarray community, the community's MGED initiative [14] created MIAME (Minimum Information About a Microarray Experiment) standard for describing microarray experiments. One drawback of this approach is that it may take a long time for a community to reach agreement on a standard and then to adopt the standard. In addition, as with mediators/wrappers it remains difficult for standards to keep up with a dynamically changing domain, especially the rapidly expanding domains of life sciences [6] [7]. Indeed, [4] further argue that any fixed interchange standard will prove to be limited as communication needs evolve.

3. A METHODOLOGY FOR SEMANTIC INTEGRATION

3.1 Proposed Methodology

This paper proposes a novel methodology to resolve and to mitigate the semantic (essential) heterogeneity problem for bioinformatics metadata based on the following proposition:

Monitoring, clustering, and appropriate visualization of metadata can help users identify patterns of practice, which in turn can facilitate the sharing and reuse of metadata, enable semantic reconciliation of existing data, and promote homogeneity among future data.

The proposed methodology focuses on clustering domain metadata in a semantically meaningful way and presenting it to the user in visually enhanced manner. Using this approach, the users (e.g., microarray researchers) can understand community trends in use of certain terms or definitions and may themselves follow these trends to share their information with the domain community and others. In this way, a dynamic standard evolution is promoted in the community. This encourages adaptation in a changing environment and can help reduce the heterogeneity of metadata from different sources.

The proposed methodology consists of the following activities:

1). *Identify candidate domain, research domain issues and locate initial, representative data sources.* Indeed, including "source, contradictory data" as recommended by [6] can improve outcomes over using only uniform data sources. In the current instantiation of the methodology, 5 publicly available microarray data sources have been selected (SMD, ArrayExpress, Genex, ArrayDB and ExpressDB); the first three implement a version of MGD and all of them claim to be compliant with the MIAME standard.

2). *Discover and extract metadata.* Working with domain experts or consultation with domain sources (such as MIAME standards in case of the microarray community) is advised to

[1] DNA microarray technique provides an efficient method for testing an individual cell/tissue sample to determine the expression levels of up to 10,000 genes.

better understand the concepts of the community and to devise initial extraction routines. The extracted metadata is stored in a data repository (and monitored so as to extract later metadata information.) The intent is to have most updated and complete picture about that domain. Research issues include investigation of mechanisms for monitoring – such as web services, subscription/publishing models, and web-crawler technology.

3). *Cluster metadata into semantically meaningful groups.* The selected metadata is first tokenized (e.g., sets of objects and attributes). The user has freedom to select some or all objects. The "ideal integration system should truly take into consideration the wishes of those who will be using the system" [6] and this user selection can additionally serve to reduce the time and interactions required to use the system.

Next, the tokenized data is sent to the clustering engine (Self-Organizing Map [9] is adapted in this approach). The clustering result may also be saved in the data repository for future use. Clustering analysis is a typical approach to organize previously unclassified datasets and is especially appropriate for exploring interrelationships among the data points [8]. With clustering, the system can potentially classify metadata into semantically related groups and present users with a high-level summary of matching result clusters rather than a long hit list. This not only improves information retrieval effectiveness [19], but also helps resolve the semantic heterogeneity of those metadata.

The clustering methods to be used for a community are developed, validated, and refined using experimentation to make sure that the clusters are meaningful to the community. Domain experts may be invited to review clustering of the metadata, or asked to provide their own metadata integration (clustering). Their results are compared to the ones generated by computer using metrics such as cluster recall, cluster precision, and an F-Measure to evaluate effectiveness [12] [19]. Further research and experimental validation can investigate the capability of the system using this accumulated and validated core domain metadata as a dynamic "reference set" for locating and mapping metadata from additional domain data sources.

4). *Visualize the clustered metadata.* The clustering result is presented to the users in two or three-dimensional views. Visualization gives the users an easy way to explore desired information from large amounts of data sources. The users should have the ability to drill down and annotate certain clusters. This is important for promoting reuse of the metadata. Labeling of resulting clusters, effectiveness of display formats to ensure reuse, and user annotation of metadata are important research issues.

3.2 MicroSEEDS Architecture

The proposed research methodology has been applied on LDAP directory metadata [10] [12]. A research architecture, Semantically Enhanced Enterprise Directory Service (SEEDS) [11], and a research prototype [25] have been developed. Experimental validation has shown that the SEEDS prototype system can successfully cluster LDAP metadata at the domain expert level. LDAP metadata tends to be reasonably well structured – with schema discovery being a core function of the protocol. Microarray experimental metadata, on the other hand, is less structured but we expect the methodology to be equally applicable and useful to the microarray domain. Working in the microarray domain will help in testing and refining the methodology to enhance its generalizability across different domains and so improving the overall toolset development.

Building on SEEDS research architecture, MicroSEEDS (Figure 1) architecture is proposed for bioinformatics metadata, specifically, microarray experimental metadata.

MicroSEEDS is composed of four components: External Information, MicroSEEDS Metadata, Semantic Homogeneity Promotion, and Metadata Access. The functionality of each component is described as follows.

- *External Information*: Existing bioinformatics sources. These sources are usually widely distributed and in different format, e.g., flat files, relational databases, or XML based repositories
- *MicroSEEDS Metadata:* Attributes (atomic metadata descriptors) and Objectclasses (consisting of a set of attributes). The terms are broadly defined so that external information can be easily transformed to MicroSEEDS data format.
- *Semantic Homogeneity Promotion***:** Metadata Updater can extract metadata from external sources (public-available or user specified) and transform them to MicroSEEDS format. Users can also use it to modify/export their metadata. The users can also subscribe the "new metadata" alerts. The Semantic Facilitator™ SM [25] is the key component of the architecture. It can interact with the users; cluster objects into semantically related groups based on a user's selection, and present results to the users using two or three-dimensional diagrams.
- *Metadata Access:* Users (microarray researchers) who want to access metadata.

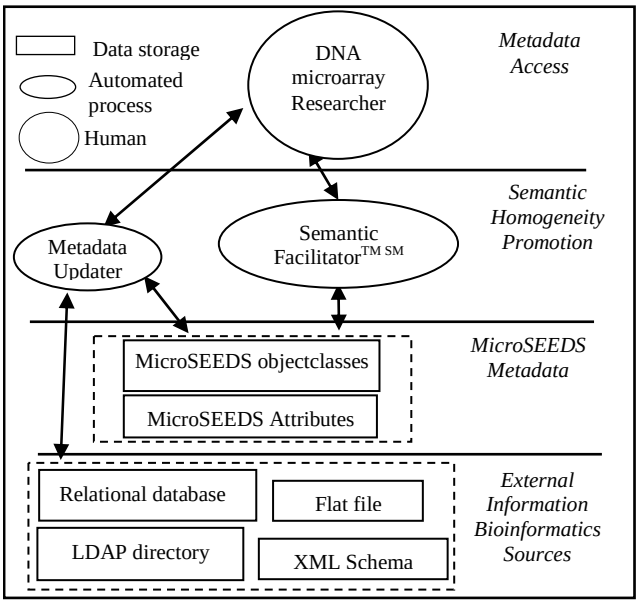

Figure 1: MicroSEEDS Architecture

3.3 MicroSEEDS Prototype Development

MicroSEEDS prototype will extend a working implementation of the SEEDS prototype designed for LDAP directory metadata

[25]. SEEDS' enabling technology components include a web-based GUI, a servlet engine (Apache/Tomcat) for business logic, a relational database repository, tokenizing and clustering routines for Self-Organizing Maps (SOM), and visualization options using SOM and the Stereoscopic Field Analyzer (SFA) [20]. SFA allows interactive manipulation of a 3D space that can improve understanding of high-dimensional data spaces [20] with its multi-dimensional nature. The Semantic Facilitator™ ℠ component helps promote semantic homogeneity by presenting metadata in related groups. With metadata seen as data to be discovered by users, the prototype uses two successful information retrieval techniques, keyword retrieval and SOM [13] [19], for discovery of metadata. The results of keyword retrieval are typically presented as a "hit list" showing matching items, but recent research shows that presenting information as higher-level groups (or clusters) of matching items can improve retrieval effectiveness [19].

The SEEDS prototype provides key enabling technology components that MicroSEEDS prototype will adopt.

3.4 Envisioned Uses of MicroSEEDS – A Scenario

This section presents a scenario illustrating how the MicroSEEDS architecture can be used to address the concerns of the microarray researcher outlined in the introduction.

A DNA microarray researcher needs to verify her microarray data experiment results against other microarray experiments. Instead of going through a potentially tedious and long process of web searching and reviewing pages related to DNA microarray among various, distributed sites, she uses MicroSEEDS functions to facilitate her search.

Locate Relevant Sources – Expected and Unexpected. The researcher logs into a MicroSEEDS' web-based interface (modeled after SEEDS) to use a Semantic Facilitator™ ℠ component to search a catalog of microarray data sources.

The MicroSEEDS data repository holds metadata of a number of microarray experiment sites/sources (these may have been added by other researchers, or perhaps by administrators of the MicroSEEDS service, or from some publicly listed sites that MGED). The researcher doesn't know which of these currently cataloged sites/sources may be helpful to her. From the catalog she selects a number of sites/sources that look promising - she deliberately includes a reference site she knows is useful to her (e.g., ArrayExpress).

By clustering the metadata of these multiple sites - including the reference site - she sees that some of the selected sites/sources have metadata more closely matching her reference site (i.e. clustered nearby the reference site). This provides her with a first level of understanding about which sites/sources may be potentially helpful.

She then engages in a computer assisted, iterative process

- She selects a site/source she is interested in from among those initially clustered.

- The interface shows her all the metadata of that one site/source.

- Default clustering of all the site's metadata helps her sees how the metadata inter-relates.

- She can then re-cluster, and by clicking on only metadata she is interested in and clustering with respect to that metadata only, she better understands the relationship of the other metadata.

With this information, the researcher can now select from among the many potential microarray experiment sites/sources the ones that she needs to study further. She may even use the selected sites she has found as a larger reference set and cluster all the sites/sources with respect to them. She has avoided looking through all sites one-by-one to locate relevant sites and instead used the MicroSEEDS tool to focus and refine her selection to sites that are most closely related to her needs. It is important to note that MicroSEEDS also clustered objects that more loosely conform to MIAME metadata (essential heterogeneity problem). The researcher would have missed these objects if she used a strict MIAME based query. By using MicroSEEDS, the researcher not only will extend the search space (to data sources less strictly conforming to MIAME and/or recently released data using approximations of MIAME formats), but also may potentially extend the completeness of the search (clustering provides relative context for semantically related objects).

Drill Down and Annotation. With the search results presented in the user interface, the researcher would point and click objects of interest, "drilling down" to display their specific metadata. The researcher may drag related objects to a design area, add an annotative comment ("new metadata"), and update the MicroSEEDS repository with "new metadata."

Publish/Subscribe Metadata and Alerts. Finally, she may upload her own microarray experimental metadata (tables, attributes...) to the MicroSEEDS, adding annotations as needed, so increasing the MicroSEEDS value to her microarray data community. The researcher optionally could subscribe to "new metadata" alerts, having MicroSEEDS report new DNA microarray sources or metadata trends.

4. DISCUSSION AND FUTURE WORK

The fast growing heterogeneous bioinformatics data sources bring big challenges to the effective and efficient use of such data. This paper describes a methodology to reconcile the semantic heterogeneity of metadata in a bioinformatics domain, specifically, microarray experiment domain. This approach can facilitate semantic reconciliation of metadata and promote semantic homogeneity through sharing and reuse of the metadata. As metadata is the key to access a data source, this research also has potential impact on *data* integration of bioinformatics data sources. The approach we propose here to deal with the metadata of microarray sources can be extended to other databases. In particular, we see a great benefit in applying this methodology to a number of data sources, particularly, mutational data such as SNP's (single nucleotide polymorphisms), protein structure databases (such as PDB or SCOP), Protein motif and domain databases (such as PROSITE, PFAM) and a variety of databases related to genetic evolution (e.g., COG or CDD).

The research architecture for MicroSEEDS applies enabling technology components of a working research prototype

(SEEDS) toward resolving metadata of LDAP directories. Despite the presence of an LDAP standard, directory vendors and users deploy directories with significant variations in metadata – both in format of metadata or the attributes and objects described – very similar to the variations found in DNA microarray data sources. SEEDS results show that such varied metadata can be extracted, tokenized and clustered into semantically related groups, comparable to human experts, and valuable to users.

Instantiating the proposed research methodology by studying the use of MicroSEEDS in DNA microarray domain seems to hold promise. Work is underway and future research tasks include: 1) metadata analysis of bioinformatics sources for DNA microarray research; 2) prototype design of MicroSEEDS working model and component implementation with focus on web-based services; 3) metrics development and evaluation of additional clustering techniques; 4) visualization development; evaluation of user interface; 5) experimental validation; and 6) web-crawling agents to acquire and catalog microarray data sources based on known metadata reference sets.

The web-based genomic, proteomic and mutational databases have not gained heavily from other web-related developments such as RDF (resource description framework, see http://www.w3.org/RDF/,) that is a way to describe the semantics of a web resource, or other ontology related developments. However, there is a move to standardize terminologies (UMLS and MeSh standards) and the major databases like Genbank and Swiss-prot are also considering incorporating web-services. Methodologies such as the one proposed above may be able to help the scientists even more in light of these developments.

5. ACKNOWLEDGEMENT

This work is partially supported by NSF ITR Grant IIS-0312636, NSF NMI Grant No. ANI-0123937, Sun Microsystems Academic Equipment Grant EDUD 7824-010460-US, Georgia State University Brain and Behavior Fellowship program, Georgia State University's Robinson College of Business, and Georgia State University's Information Systems & Technology.

6. REFERENCE

[1] Batini, C., Lenzerini, M. and Navathe, S.B. A Comparative Analysis of Methodologies for Database Schema Integration. *ACM Computing Surveys*, 18, 4, December 1986, pp. 323-364.

[2] Brooks, F. P. No Silver Bullet: Essence and Accidents of Software Engineering. *Computer*, 20 (4), pp. 10-19, 1987.

[3] Chen, L., Jamil, H. M., and Wang, N. Automatic Composite Wrapper Generation for Semi-Structured Biological Data Based on Table Structure Identification. *SIGMOD Record* 33(2): 58-64, 2004.

[4] Damsgaard, J.and Truex, D. Binary Trading Relations and the Limits of EDI Standards: The Procrustean Bed of Standards. *European Journal of Information Systems*, 9 (3), pp. 173-188, 2000.

[5] Foster, I. and Grossman, R. L. Data Integration in a Bandwidth-rich World. *Communications of the ACM*, vol. 6, no. 11, November 2003. pp50-57.

[6] Hernandez, T. and Kambhampati, S. Integration of Biological Sources: Current Systems and Challenges Ahead. *SIGMOD Record* 33(3): 51-60 (2004)

[7] Jagadish, H. V. and Olken, F. Database Management for Life Science Research: Summary Report of the Workshop on Data Management for Molecular and Cell Biology at the National Library of Medicine, Bethesda, Maryland, February 2-3, 2003. *OMICS A Journal of Integrative Biology*, 7 (1), 2003.

[8] Jain, A. K., Murty, M. N., and Flynn, P. J. Data Clustering: A Review. *ACM Computing Surveys*, 31, 3, pp. 264-323, 1999.

[9] Kohonen, T. *Self-Organizing Maps*, Springer-Verlag, Berlin, 1995.

[10] Kuechler, D., Vaishnavi, V., and. Vandenberg, A. An Architecture to Support Communities of Interest Using Directory Services Capabilities. *Proceedings Hawaii International Conference on System Sciences*, Big Island, Hawaii, 2003.

[11] Li, L., Vaishnavi, V., and Vandenberg, A. An Architecture for Semantic Facilitation and Reuse of Directory Metadata. *Proc. 2004 International Conference on Information and Knowledge Engineering*, Las Vegas, Nevada, 2004.

[12] Liang, J., Vaishnavi, V., and Vandenberg, A. Clustering of LDAP Directory Schemas to Facilitate Information Resources Interoperability Across Organizations. *IEEE Transactions on Systems, Man, and Cybernetics, Part A* (to appear).

[13] Liu, Y., Ciliax, B. J., Borges, K., Dasigi, V., Ram, A., Navathe, S., and Dingledine, R. Comparison of Two Schemes for Automatic Keyword Extraction from MEDLINE for Functional Gene Clustering. *IEEE Conf. on Computational Systems Bioinformatics* (CSB'2004), August 2004.

[14] MGED: Microarray Gene Expression Data Society (MGED), home page, 2004. <http://www.mged.org> (Last accessed November 14, 2004).

[15]. Navathe, S. and Patil, U. Genomic and Proteomic Databases and Applications: A challenge for Database Technology," *Proc. 9th International Conference on Database Systems for Advanced Applications* (DASFAA 2004), Jeju Island, Korea, March 2004, - Invited Paper.

[16] Newman, H. B., Ellisman, M. H., and Orcutt J. A. Data-Intensive E-Science Frontier. *Communications of the ACM*, 46 (11), pp. 68-77, November 2003.

[17] Panayiota, P. and Nicholas, D. Familiarity with and Use of Metadata Formats and Metadata Registries amongst Those Working in Diverse Professional Communities within the Information Sector. *Aslib Proceedings*, 53, 8, pp. 309-324, 2001

[18] Roszkiewicz, R. Metadata in Context. *The Seybold Repor*t, vol. 4, no. 8, 2004.

[19] Roussinov, D. Information Foraging Through Clustering and Summarization: A Self-Organizing Approach. A Dissertation Submitted to the Faculty of the Committee on Business Administration, the University of Arizona, 1999.

[20] Shaw, C. D., Hall, J. A., Ebertc D. S., and Roberts, D. A., "Interactive Lens Visualization Techniques," in *IEEE Visualization'99*, pp. 155-160, October 1999.

[21] Sheth, A., Gala, S. K., Navathe, S. B. On Automatic Reasoning for Schema Integration. *Int. Journal of Intelligent Co-operative Information Systems*, 2 (1), March 1993.

[22] Stekel, D., *Microarray Bioinformatics*, Cambridge University Press, 2003.

[23] Stoimenov, L., Djordjevic, K., Stojanovic, D. Integration of GIS Data Sources over the Internet Using Mediator and Wrapper Technology. *Proceedings of the 2000 10th Mediterranean Electrotechnical Conference. Information Technology and Electrotechnology for the Mediterranean Countries (MeleCon 2000)*, pp. 334-336, 2000.

[24] Vaishnavi, V. and Kuechler, W. Universal Enterprise Integration: The Challenges of and Approaches to Web-Enabled Virtual Organizations. *Information Technology & Management*, 6 (1), 2005, to appear.

[25] Vandenberg, A., Liang, J., Bolet, V., Kou, H., Vaishnavi, V., and Kuechler, D. Research Prototype: Semantic Facilitator ™ SM for LDAP Directory Services. *Proceedings of the 12th Annual Workshop on Information Technologies and Systems*, 2002.