# *Ontology-based information integration using INDUS system*

*Doina Caragea', Jie Bao, Jyotishman Pathak and Vasant Honavar*

*Artificial Intelligence Research Laboratory, Department of Computer Science, Iowa State University, Ames, IA 50010, USA*

**ABSTRACT**

INDUS (Intelligent Data Understanding System) is a *federated*, *query-centric* system for information integration and knowledge acquisition from distributed semantically heterogeneous data sources. INDUS employs ontologies (controlled vocabularies of domain specific terms, and relationships among terms) and inter-ontology mappings, to enable a user to view a collection of such data sources (regardless of location, internal structure and query interfaces) as though they were a collection of tables structured according to a user-supplied ontology.

## 1 INTRODUCTION

Ongoing transformation of biology from a data-poor science into an increasingly data-rich science has resulted in a large number of autonomous data sources (e.g., repositories of protein sequences, structures, expression patterns, interactions). This has led to unprecedented, and as yet, largely unrealized opportunities for large-scale collaborative discovery in a number of areas: characterization of macromolecular sequence-structure-function relationships, discovery of complex genetic regulatory networks, among others.

At present, there are hundreds of databases of interest to molecular biologists alone [Discala et al., 2000]. Because the data repositories are typically autonomous, and often focused on specific subfields of biology, ontological (and hence semantic) differences among them are simply unavoidable. However, in exploring specific scientific questions of interest, scientists often need to be able to retrieve and analyze data from multiple sources. Effective use of such data in a given context requires reconciliation of semantic differences among the relevant data sources from a user's point of view. Hence, there is an urgent need for tools to support rapid and flexible assembly and analysis of data from semantically heterogeneous data sources [Jagadish and Olken, 2003].

## 2 APPROACH

INDUS is a *federated*, *query-centric* system for data integration and knowledge acquisition from distributed, semantically heterogeneous data (See **Fig. 1**). INDUS makes explicit data source specific information, such as the data source schema and (the typically implicit) data source ontologies. The resulting ontology-extended data sources [Caragea et al., 2004] enable users to specify semantic correspondences between the user ontology and the data source ontologies by specifying inter-ontology mappings.

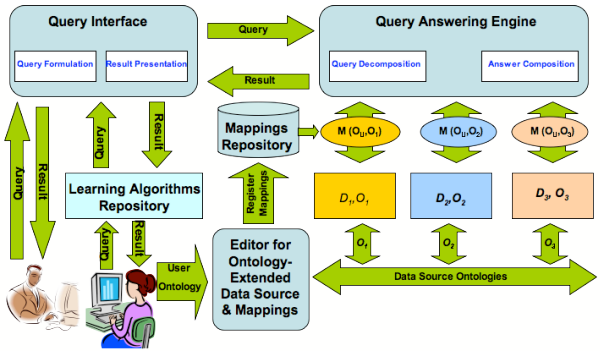


**Fig. 1.** INDUS: a system for data integration and knowledge acquisition from semantically heterogeneous distributed data.

This enables each user to view a collection of autonomous, semantically heterogeneous, distributed data as though they were a collection of inter-related tables structured according to an individual user's ontology. Thus, users can interact with and explore data sources of interest to them from multiple points of view simply by changing their perspective (i.e., user ontology and semantic correspondences between the user ontology and the data source ontologies). Queries posed using terms in the user ontology are transformed, using a sound query rewriting algorithm, into queries that can be answered by the individual data sources. The results are expressed in terms of the user's ontology [Caragea et al., 2004] (See **Fig. 1**).

## 3 INDUS PROTOTYPE

We have completed the implementation of a working prototype of the INDUS system to enable biologists with some familiarity with the relevant data sources to rapidly and flexibly assemble data sets from multiple data sources and to query these data sets. This can be done by specifying a user ontology, simple semantic mappings between data source specific ontologies and the user ontology and queries – all without having to write any code. An initial version of the INDUS software and documentation are available at www.cild.iastate.edu/GM066387_homepage.htm.

* To whom correspondence should be addressed.

The current implementation includes support for:

- Import and reuse of selected fragments of existing ontologies (e.g., Gene Ontology GO), editing of ontologies, specification of semantic relationships between ontologies using inter-ontology mappings [Bao and Honavar, 2004].
- Specification of semantic correspondences between a user ontology $O_U$ and data source ontologies $O_D$ [Caragea et al., 2004]. Semantic correspondences between ontologies can be defined at two levels: schema level (between attributes that define data source schemas) and attribute level (between values of attributes). INDUS allows the following types of semantic correspondences at both schema and attribute level: semantic equality (e.g., *AASequence:*$O_D \equiv$ *ProteinSequence:*$O_U$), semantic subsumption (e.g., *MIPS:16.19.01:*$O_D \leq$ *GO: 0017076:*$O_U$) and procedural mappings (e.g., from *AASequence:*$O_D$ to *AAComposition:*$O_U$). Consistency of semantic correspondences are verified by an efficient algorithm for reasoning about subsumption and equivalence relationships.
- Registration of a new data source using a data-source editor for defining the schema of the data source (the names of the attributes and their corresponding ontological types), location, type of the data source and access procedures that can be used to interact with a data source. In the current implementation several types of data sources can be defined including multiple relational databases (Oracle, MySQL, PostgreSQL), and files (e.g., ARFF files used in WEKA, a widely used open source machine learning software package).
- Specification and execution of queries across multiple large, semantically heterogeneous data sources with different interfaces, functionalities and access restrictions. Each user may choose relevant data sources from a list of data sources that have been previously registered with the system and specify a user ontology (by selecting an ontology from a list of available ontologies or by invoking the ontology editor and defining a new ontology).

Once the ontology-extended data sources and the user ontology have been specified, the user can select mappings between data source ontologies and user ontology from the available set of existent mappings (or invoke the mappings editor to define a new set of mappings). Once the necessary mappings are specified, the system can answer queries posed by the user. The data needed for answering a query is specified by selecting (and possibly restricting) attributes from the user ontology, through a user-friendly interface. Queries posed by the user are sent to a query-answering engine (QAE) that decomposes a user query into subqueries that can be answered by the individual data sources (using predefined or user-supplied mappings between the respective ontologies). The answer to the user query (expressed in terms of user ontology) is constructed and presented to the user by the QAE using results of queries to the distributed data sources. INDUS has been used to assemble several data sets used in the exploration of protein sequence-structure-function relationships [Caragea et al., 2005]. Examples of such data sets include: a data set used for building a classifier for automating functional annotation of protein sequences based on sequence composition [Andorf et al., 2004] and structural features of proteins and a comprehensive database of protein-protein interfaces www.cild.iastate.edu/GM066387_homepage.htm.

## 4 CONCLUSIONS

We have presented INDUS, a federated, query-centric approach to answering user queries from distributed, semantically heterogeneous data sources. INDUS assumes a clear separation between data and the semantics of the data (ontologies) and allows users to specify ontologies and mappings between data source ontologies and user ontology. INDUS enables users (or application programs e.g., learning algorithms) to retrieve results of queries from semantically heterogeneous data sources.

## ACKNOWLEDGEMENTS

This work was funded in part by grants from the National Science Foundation (IIS 0219699) and the National Institutes of Health (GM 066387).

## REFERENCES

Andorf, C., Silvescu, A., Dobbs, D. and Honavar, V. (2004). Learning Classifiers for Assigning Protein Sequences to Gene Ontology Functional Families. In: *Fifth International Conference on Knowledge Based Computer Systems (KBCS 2004)*, India.

Bao, J. and Honavar, V. (2004). Collaborative ontology building with wiki@nt - a multi-agent based ontology building environment. In: *Proc. of 3rd International Workshop on Evaluation of Ontology based Tools*, ISWC 2004, Japan.

Caragea, D., Pathak, J., and Honavar, V. (2004). Learning Classifiers from Semantically Heterogeneous Data. In: *Proceedings of the Third International Conference on Ontologies, DataBases and Applications of Semantics for Large Scale Information Systems* (ODBASE'04), October 25-29, 2004, Agia Napa, Cyprus.

Caragea, D., Silvescu, A., Pathak, J., Bao, J., Andorf., C., Dobbs, D. and Honavar, V. (2005). *Information Integration and Knowledge Acquisition from Semantically Heterogeneous Biological Data Sources*. In: Proc. of the 2nd Int. Workshop on Data Integration in Life Sciences (DILS'05), San Diego, CA.

Discala, C., Benigni, X. Barillot, E. and Vaysseix, G. (2000). DBcat: a catalog of 500 biological databases. Nucleic Acids Res. 2000 Jan 1;28(1):8-9.

Jagadish, H.V. and Olken, F. (2003). Data Management for the Biosciences. Report of the NSF/NLM Workshop of Data Management for Molecular and Cell Biology, Feb. 2-3, 2003.