



Fall 2008 CIS Distinguished Lecture Series

Orchestra: Sharing Inconsistent Data in a Consistent Way

Zachary Ives
University of Pennsylvania
October 1, 2008

Abstract: One of the most pressing needs in business, government, and science is to bring together structured data from a variety of systems, formats, and terminologies. For instance, the emerging field of systems biology seeks to unify biological data to get a big-picture view of the processes within living organisms. Many organizations have set up databases designed to be "clearing houses" for specific types of information: each is separately maintained, cleaned, and curated, and has its own schema and terminology. Updates are constantly made as hypothesized relationships are confirmed or refuted, or new discoveries are made. The different databases contain complementary information that must be integrated to get a complete picture - and each database may have data of different quality or relevance to a domain. However, there is often no consensus on what the definitive answers are - each site may have different beliefs.

The Orchestra project focuses on how to support exchange of data (and updates) among collaborators with evolving databases, in a way that accommodates disagreement, different schemas, and different levels of authority and quality. Orchestra considers collaborators' databases to be logical *peers* into which data can be imported and then locally modified. It allows for a network of *schema mappings* that interrelate peers, annotated with *trust policies* specifying the conditions under which a peer is willing to import data. As a data item is mapped from site to site in the system, its *provenance* is recorded; a peer's trust policies use this provenance (and the values of the data) to assign a score to each incoming data item (based on perceived quality or relevance), and the peer then uses this score to reconcile conflicts and compute a consistent data instance, whose contents may be unique to the peer. The scores assigned to the individual sources can even be *learned* based on user feedback about query answers. The end result is a system that allows each database to selectively diverge from the others as appropriate, but to remain "in sync" in all other cases.

Bio: Zachary Ives is an Assistant Professor at the University of Pennsylvania and an Associated Faculty Member of the Penn Center for Bioinformatics. He received his B.S. from Sonoma State University and his PhD from the University of Washington. His research interests include data integration, peer-to-peer models of data sharing, processing and security of heterogeneous sensor streams, and data exchange between autonomous systems. He is a recipient of the NSF CAREER award and a member of the 2006 (first) DARPA Computer Science Study Panel. He has been a co-program chair for the XML Symposium (2006) and New Trends in Information Integration (2008) workshops.

Location: 4th Floor Conference Room (Wachman 447)

Time: 3-4pm, Wednesday, October 1, 2008

Refreshments will be served!