

SHARQ Guide: Finding relevant biological data and queries in a peer data management system

Sarah Cohen-Boulakia, Olivier Biton, Shirley Cohen, Zachary Ives, Val Tannen and
Susan Davidson

Department of Computer and Information Science
University of Pennsylvania, Philadelphia, USA

Scientists are now faced with an explosion of information which must be rapidly analyzed to form hypotheses and create knowledge. A major challenge lies in how to effectively share information among collaborating, yet autonomous, biological partners. These partners are peers in the sense that they have a diversity of perspectives (and hence heterogeneous and complementary schemas), dynamic data, and the possibility of intermittent connectivity or participation.

SHARQ (Sharing Heterogeneous and Autonomous Resources and Queries) is a collaborative project between the database group at the University of Pennsylvania (Penn) and two biological research groups from Penn Center for Bioinformatics and the Children's Hospital of Philadelphia. The goal of SHARQ is to develop generic tools and technologies for creating and maintaining confederations of peers whose purpose is distributed data sharing. In response to the difficulties outlined, our solution will emphasize: (i) decentralization achieving both scalability and flexibility, (ii) incremental development of resources such as schemas, mappings between different schemas, and queries, (iii) rapid discovery mechanisms for finding the resources relevant to a topic, and (iv) tolerance for intermittent participation of members and for approximate consistency of mappings.

The core engine of SHARQ is the Orchestra system (Ives et al, 2002 & 2005, Taylor et al, 2006, Green et al, 2006), which builds upon concepts from the Piazza peer data management system (PDMS) (Halevy et al, 2004 & 2005). Orchestra supports the exchange of data and updates among cooperating, heterogeneous databases, making use of policies to quickly and automatically manage disagreement among conflicting data. However, knowing what information is available in the peer network is difficult to determine. The SHARQ Guide is therefore being designed to enable biologists to find relevant information within a peer data management system. It provides assistance not only for users who ask queries, but also for owners of peers who wish to be registered within the Guide. Key ideas of the SHARQ Guide include: (i) representing biological entities and relationships as a graph, following the approach of BioGuide (Cohen-Boulakia et al, 2005). This graph can be extended in a collaborative way by the peer administrators. (ii) Helping administrators of peers to register their schema. (iii) Expressing queries without having to know/cite the schemas to use for querying (transparent queries). (iii) Proposing new features to maximize the amount of data returned to the user, by allowing some fields in the query to be optional.

More information about SHARQ and about the papers cited above is available at
<http://db.cis.upenn.edu/research/SHARQ.html>.