# Schema Mapping and Data Integration with Clio

Barbara Eckman, Mauricio Hernández, Howard Ho, Felix Naumann, Lucian Popa
IBM Life Sciences Solutions & IBM Almaden Research Center
{baeckman,mauricio,ho,felix,lucian}@us.ibm.com

Bioinformatics data sources typically have large, complex structures, reflecting the richness of the scientific concepts they model. These structures can be difficult to understand for those who are not accustomed to them. Many bioinformatics data sources cover roughly the same domain, such as genes, proteins, sequence annotations, or microarray results. To derive the greatest benefit for scientific investigation, it is necessary to provide an integrated view of all related data sources.

The difficulties of such integration are twofold: (1) Data from different sources is typically structured differently, i.e., it conforms to different schemas, and it is necessary to understand each of the different schemas to make effective use of the data. (2) Sources often overlap in the data they cover, i.e., they store data about the same entities but possibly with conflicting data values. There is a critical need for a user-friendly tool to transform data from one database schema to another, and to discover corresponding data in two different databases, regardless of the structure of the databases or the names that are given to corresponding attributes.

Clio is an information integration tool that helps meet both these needs. Clio semi-automatically defines a mapping from one or more source schemas to a target schema, and generates a set of queries that transform and integrate data from those sources to conform to the target schema. Such queries can be used to populate data warehouses, but also to define views in a non-materialized, federated integration environment like DiscoveryLink. Sources and target can be any combination of relational databases (such as DB2 UDB, Sybase, or Oracle) and XML data.

Several components in Clio support users in finding such a mapping: The user-friendly *schema-viewer* allows drawing arrows between source schema elements and corresponding schema elements in the target. Such arrows may cross nesting levels, combine multiple elements, split and merge tables, etc. Clio incrementally interprets these arrows as mappings and generates correct queries accordingly. Because users can get lost in very large and unfamiliar schemas, an *attribute-matcher* component automatically suggests likely mappings by analyzing the schemas and the underlying data. For each attribute, Clio extracts features from small, random database samples. Using these features, a Naïve Bayes-based classifier finds similar attributes and suggests mappings between them. The final result of a mapping is a set of queries that take data from the sources and produce data conforming to the target schema. Depending on the source type, the queries are formulated in SQL, XQuery, or XSLT.

A specialized application of Clio reads a nested source schema, for instance the SwissProt XML schema, and automatically generates a relational target schema, according to one of several "shredding" strategies. Using this relational target, Clio generates relational DDL statements, which enable the user either to create a relational warehouse into which the data may be imported, or to issue SQL queries directly against remote XML documents or XML WebServices through DiscoveryLink, IBM's federated data integration system. For more information see http://www.almaden.ibm.com/cs/clio/ .