Statement of interest: inconsistency-tolerance in data integration systems

Riccardo Rosati
Dipartimento di Informatica e Sistemistica
Sapienza Università di Roma, Italy

The task of a data integration system is to combine the data residing at different, autonomous sources, and providing the user with a unified view of these data, called *global schema*. Users query the global schema, while the system carries out the task of suitably accessing different sources and assembling the data retrieved at each source into the final answer to the query.

Since sources are in general autonomous subsystems, the information provided by the data at the sources are likely not to be consistent with the knowledge (constraints) expressed by the global schema.

Current data integration technology is actually unable to handle sources that are inconsistent with the global schema: in fact, data integration systems mainly deal with this problem through a (static) data cleaning approach, i.e., data that has to be integrated is modified in order to recover consistency with respect to the global schema. However, in many situations it would be much more desirable to derive significant information from the database even in the presence of data inconsistent with the global schema. Indeed, in many application scenarios, the explicit repair of data is not convenient, or even impossible: e.g., in virtual data integration, sources are not controlled by the integration system, which in general is not allowed to modify source data; moreover, data are not materialized (copied) in the data integration system, thus they cannot be modified.

On the other hand, following the ideas proposed by the research in *consistent query answering in databases*, it might be possible to develop query answering (and more generally data management) techniques that realize a dynamic, *virtual* repair of data. According to such an approach, data are not cleaned, and inconsistency is handled at query evaluation time, through suitable query answering methods which are able to extract significant information from a data integration system even in the presence of inconsistent data.

Recent research in inconsistency-tolerance in databases and information systems has produced interesting results, and promising techniques (some of which are based on standard relational database technology) have been proposed. So, it would be interesting to discuss whether (and how) this kind of technology may now have a potential impact on commercial data integration systems and applications. In particular:

- from the application/industrial side, is there a real interest for the development of technologies for inconsistency-tolerance in data integration and data exchange?
- what are the forms of inconsistency-tolerance that are more interesting for the current data integration and data exchange applications?
- how far is research from the development of *effective* methods and techniques for inconsistency-tolerance in data integration?