

29

Issues in Schema Integration: Perspective of an Industrial Researcher*

Amit P. Sheth
Bellcore
444 Hoes Lane, RRC 1J-210,
Piscataway, NJ 08854 USA.
e-mail: amit@ctt.bellcore.com

I will discuss the issues of schema integration from the perspective of a researcher involved in developing and trying to apply methodology and techniques to real industrial problems in this area.

Perhaps the biggest hurdle I had to face was to modify my perception as a researcher of "what is integration and why are we doing it" to accommodate the needs of the users/practitioners. Very often, the standard notions and assumptions with which researchers have defined the methodologies and techniques fail to address the problems of real users, since the goal and assumptions of "integration" to the users are quite different. Which of the following is (are) the **goal(s) of integration?**:

- (a) To develop a global schema, access to which will give transparency to the distributed data? Are the schemas being integrated those of preexisting databases? Can these databases be changed? Is there a need for single global schema or multiple federated schemas?
- (b) To integrate views/subschemas of the same application to develop the database schema for that application?
- (c) To analyze schemas for dependencies among them to facilitate interoperability? (I differentiate between integration, which implies generation of new objects by merging existing objects and generation of a mapping for processing queries given on integrated objects, and dependency analysis and specification, which identifies relationship between related objects but does not create new objects.)
- (d) To develop (the schema of) a new application that encompasses one or more existing applications?
- (e) To perform a *shallow* analysis/integration (e.g., without attributes, functions, or behavior), and/or *deep* analysis/integration resulting in complete dependency/integration mapping?
- (f) To address also the issue of updates and data consistency? Most research on view/schema integration has not addressed this issue.

Different goals for integration and dependency analysis result in altogether different expectations from the methodology, techniques, and tools. I feel that research to date has largely focused on the first two of the above objectives, but ignored the last four.

Developing a complete solution to meet one or more of the goals requires investigating alternative methodologies, developing techniques, and developing tools to support the selected methodology and techniques. Issues to consider about **methodology**

*This abstract is an extended version of the position statement given at panel on "Cooperative Database Design" at VLDB'91, Barcelona, Spain, September 3-6, 1991.

are:

- (a) We cannot assume that the views we start with are either complete, correct, or consistent. How do we accommodate additional information supplied during the integration process and the corrections made to the schemas/views we start with?
- (b) Will multiple users cooperate in the analysis/integration process? Is cooperative design addressing the user interface issue to allow multiple designers/users/DBAs to simultaneously work on a *single* analysis/integration process? Or is it something more involved, for example, to accommodate and/or analyze multiple integration options identified/developed by concurrently working designers? Personally, I see many problems to be solved before the use of cooperative design and integration is of significant value for the potential users I have interacted with.
- (c) Does the methodology support top-down, bottom-up, or mixed process? Does it support one or more of the goals of integration described above?

Issues to consider about **techniques** are:

- (a) Is the main purpose that of mechanizing and book keeping of a current manual process? Is it to relieve the designers of some of the decision making process?
- (b) Are there language or naming standards that can help in concept comparison? I see integration as ultimately a process driven by semantics. What is the role of structure and name comparison?

The techniques for schema/view analysis or concept comparison are the key to all of the above goals if some of the automated reasoning has to be supported. My opinion is that to support the multiple goals identified earlier, no single technique is sufficient (or has been proven so). That is why I advocate a toolkit approach that can provide multiple techniques to achieve various (sub)goals of database design and integration.

At Bellcore we have developed a prototype software toolkit called *BERDI* (*Bellcore E-R Schema Design and Integration Toolkit*). We incrementally add techniques to the toolkit and test the methodology and techniques that the toolkit supports by applying them to real schemas¹.

A partial taxonomy of techniques to support comparison of concepts/objects consists of the following categories.

- (a) Graphical facilities and query languages, CASE techniques/tools
- (b) Formal/logic-based techniques: classification (e.g., based on terminological logic), logic based approach, constraint analysis, model-based specifications, structural integration
- (c) AI/heuristic techniques: expert systems, case based reasoning, learning (on schema and/or instances)
- (d) Formal language, natural language
- (e) Use of a large existing knowledge base (e.g., Cyc), use of thesaurus/dictionary/metadata

The categories of this taxonomy are not completely disjoint- they overlap in some cases and complement in some other. The taxonomy shows that techniques from many different branches of computer science are being investigated to address the interesting but

¹ *BERDI* is not an alternative to a production CASE tool, but a prototype to test and demonstrate methodology and techniques related to schema analysis and integration. The work at Bellcore is jointly performed with Howard Marcus.

difficult problem of *semantic equivalence* (concept/attribute equivalence/relationship) that is at the heart of any integration effort.

At Bellcore, our goal is to support all of the analysis and integration goals described earlier and support the multiple promising techniques. *BERDI* currently supports, to various degrees of satisfaction and completeness, all but the last of the goals of analysis and integration. It supports graphical interface for querying, analyzing, and integrating schemas, and a classification based automatic reasoning² to identify the relationships among object classes (or entity types). Our work on structural integration using object-oriented *dual mode*³ has not yet been implemented.

Current focus of our work on schema integration is to understand whether our methodology and the techniques are useful in integrating real or realistic schemas and views. This work is not yet complete; however, the researchers among us may find some general observations on how realistic schemas compare with those used in most research efforts (e.g., data model issues, how a schema relates to the database it models, size of the schemas, structures/properties of the schemas) interesting. While this will be far from developing a "benchmark" that can be used to compare the techniques, it could help us in determining whether the techniques we are investigating might interest practitioners.

² Joint work with the University of Florida (Prof. Navathe, Gala, Savasere).

³ Joint work with the New Jersey Institute of Technology (Prof. Perl, Prof. Geller).