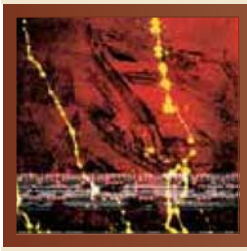
Database Conceptual Schema Matching

Marco A. Casanova, Karin K. Breitman, Daniela F. Brauner, and André L. A. Marins
Pontifícia Universidade Católica do Rio de Janeiro



Emphasis on the a priori design of standard database conceptual schemas simplifies schema matching.

A *database conceptual schema* is a high-level description of how database concepts are organized, which is typically as classes of objects and their attributes. A fundamental operation in many database applications, *schema matching* involves finding a mapping μ between the concepts in a source scheme S and the concepts in a target schema T such that, if $t = \mu(s)$, then s and t have the same meaning.

Along with data warehousing, *query mediation* relies heavily on schema matching. This application uses a *mediator* to translate user queries, formulated in terms of a common schema M, into queries that local databases can handle. The mediator must therefore be able to match each local schema with M. Query mediation is particularly challenging in the context of the Web, where the number of local databases, over which the mediator has little control, is enormous.

We examine three major approaches to schema matching—*syntactic, semantic,* and *a priori*—using examples, with a focus on mediator design.

SYNTACTIC APPROACH

This approach involves matching two schemas based on syntactical hints, such as attribute data types and naming similarities. It assumes that syntactical proximity implies semantic similarity, but such assumptions are often unwarranted and can lead to incorrect mappings.

For example, consider two schemas S and T that describe databases whose application domains aren't entirely clear. Assume that S has a set of objects named Games, with attributes Name and ESRB (Entertainment Software Rating Board), and T has a set of objects named Gaming, with attributes Name, Price, and Rating, as shown in Figure 1a. Using only syntactical similarity, Games would probably match with Gaming, and the Name attributes would definitely match with each other, but ESRB would not match with Rating.

If S and T describe computer game databases, this matching is reasonable, though it still misses the match between ESRB and Rating, which likely refers to ratings assigned by the ESRB. However, if S describes the database of a travel agency specializing in safaris, matching Games (meaning big game hunting) with Gaming (meaning computer games) is obviously inaccurate.

SEMANTIC APPROACH

This approach uses semantic clues to generate hypotheses about schema matching. It generally tries to detect how the same real-world objects are represented in two different databases and leverages this information to match the schemas.

The semantic approach is more robust than the syntactic one, but it seems to apply only when the schemas to be matched are simple. It also depends on the ability to determine whether two objects from different databases are equivalent, that is, whether they represent the same real-world instance.

Returning to the example of the schemas S and T with objects named Games and Gaming, respectively, the mediator might implement the procedure shown in Figure 1b. It could select a few typical objects stored in Gaming, such as *Flight Simulator* and *Super Mario Bros*, and probe S to check whether they indeed occur in Games. The mediator could then use this information to match Games with Gaming, the Name attributes, and ESRB with Rating, as expected.

Note that this procedure would succeed when S and T both describe computer game databases, but not when S describes a travel agency database and T describes a computer game database. By overgeneralizing, the mediator might in fact completely ignore that Games and Gaming are syntactically similar and try to match S and T only using sets of typical objects.

A PRIORI APPROACH

Both the syntactic and semantic approaches seek to detect equivalent objects stored in different databases. In doing so, they assume that there is an automatic way to detect object equivalence. However, this assumption can lead to undesirable consequences.

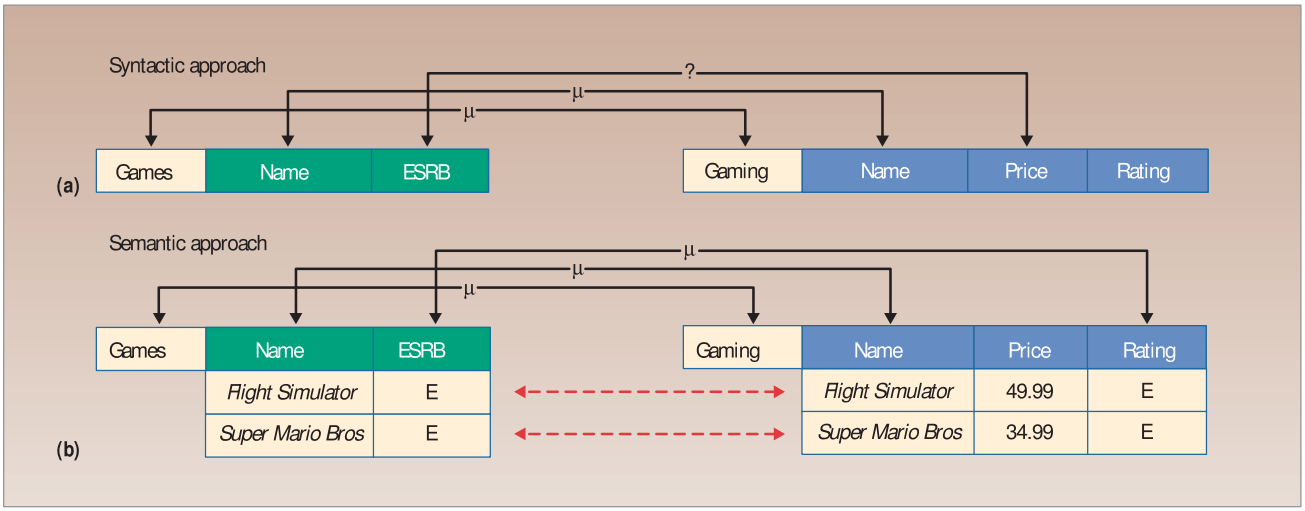


Figure 1. Examples of a posteriori schema matching: (a) syntactic approach, and (b) semantic approach.

For example, *Flight Simulator* might denote the same computer game in S and T as it is a trademark name, but matching them isn't entirely foolproof because this game has multiple versions. The mediator could use product ID numbers instead, but even this is problematic: Games sold in the US are identified by the 12-digit Universal Product Code, while games sold in Europe use the 14-digit Global Trade Item Number.

A priori versus a posteriori

In contrast to the syntactic and the semantic approaches, which are a posteriori in the sense that they try to match the schemas of preexisting databases, we propose a third alternative that we call the a priori approach.

When specifying databases that will interact with each other, the designer should first select an appropriate standard, if one exists, to guide design of the exported schemas. If none exists, the designer should publish a proposal for a common schema covering the application domain.

If both S and T follow the same common schema, matching them becomes trivial: Define the mapping μ from the concepts in S into the concepts of T in such a way that $t = \mu(s)$ if s and t denote the same concept from the common schema. Note that both S and T could actually be subsets of the common schema.

Standards

There are numerous standards that database designers can't ignore when specifying databases and publishing their content. For example, international standard ISO 19115:2003, *Geographic Information—Metadata,* defines a metadata schema to describe geographic objects, which in fact is a standard for exported schemas for geographic catalogs.

The metadata schema has a set of core elements and a set of optional elements. Each application must define a profile, that is, the list of optional elements it implements. Therefore, geographic catalogs that follow this standard can interoperate by simply exchanging their profile. The standard fixes the semantics of the exported schemas a priori.

Ontologies

As defined by the World Wide Web Consortium, ontologies can help to transform the a priori approach into a viable strategy. A reasonable strategy to define a common schema for an application domain is to

- select fragments of known, popular ontologies such as WordNet that cover the concepts pertaining to the application domain;
- align concepts from distinct fragments into unified concepts; and
- publish the unified concepts as an ontology, indicating which are mandatory and which are optional.

The published ontology should retain the origin of the unified concepts for semantic clarity.

Example

Consider the problem of designing a federation of databases that store the provenance of works of art, that is, a record of their passage through various owners.

The designer could start by observing that several upper-level ontologies such as OpenCyc, the Suggested Upper Merged Ontology (SUMO), the Common Semantic Model (COSMO), and the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) include the concepts of *event, agent,* and *action,* as Table 1 shows. Indeed, provenance can be modeled as a sequence of events that result from actions such as sell, borrow, and exhibit carried out by agents such as artists, collectors, and museums.

From these concepts, the designer could propose a common schema that any database belonging to the federation can follow to publish data describing the provenance of works of art.

The designer also could simply adopt a fragment of the draft international standard ISO 21127:2006, *Informa-*

Table 1. Suggested partial alignment of upper-level ontologies and ISO standards.

Provenance concept	OpenCyc	SUMO	COSMO	DOLCE	ISO 21127:2006	ISO 14721:2003
Event	Event	Process	Event	Event	Event	Event
Agent	Agent-generic	SentientAgent	Agent	Agent	Actor	Responsible agency
Action	Action	IntentionalProcess	Action	Action	Activity	Procedure

tion and Documentation—A Reference Ontology for the Interchange of Cultural Heritage Information, or of ISO 14721:2003, *Space Data and Information Transfer Systems—Open Archival Information System—Reference Model,* which also cover provenance.

If carefully applied, the a priori approach can resolve the intractable problem of database conceptual schema matching. The technology is readily available; all that is required is disciplined schema design.

Marco A. Casanova is a professor in the Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio). Contact him at casanova@inf.puc-rio.br.

Karin K. Breitman is a research associate in the Departamento de Informática, PUC-Rio. Contact her at karin@inf.puc-rio.br.

Daniela F. Brauner is a PhD student in the Departamento de Informática, PUC-Rio. Contact her at dani@inf.puc-rio.br.

André L.A. Marins is an MSc student in the Departamento de Informática, PUC-Rio. Contact him at amarins@inf.puc-rio.br.

Editor: Michael G. Hinchey, Loyola College in Maryland; mhinchey@loyola.edu