


Data Models for Data Integration



Oscar Romero

Facultat d'Informàtica de Barcelona

Universitat Politècnica de Catalunya

Data Modeling

- Data Model: *“Description of real-world objects, their properties and relationships”*
- According to their level of abstraction...
 - Conceptual Data Models
 - Logical Data Models
 - Database Models [Codd]
 - Data Structure, Integrity Constraints, Algebra
 - Physical Data Models

The Relevance of Data Modeling

- In a data integration system we have:
 - The data model of each data source
 - The global schema
- The global schema is a homogenised view resulting from integrating the source schemata
 - Must overcome:
 - Syntactic heterogeneities
 - Semantic heterogeneities
- Throughout this homogenisation process a canonical model is needed
 - Data sources are expressed in the canonical model
 - Then, the exposed data must be identified (export schema)
 - The global schema derives from integrating the export schemata
 - Ideally, the global schema should be expressed in terms of the canonical model

Canonical Data Models

- When several data sources are to interoperate a *common model* needs to be chosen. In data integration, this is known as the *canonical data model*
- The suitability of data models as canonical models can be analysed from the point of view of its:
 - **Expressiveness**
 - **Semantic relativeness**

Expressiveness

- Definition: The degree to which the model can directly represent (express in a natural way) a conceptualisation, no matter how complex this conceptualisation might be, and which concepts compose it
- Expressiveness can be divided into *structural* and *behavioural expressiveness*
 - Structural expressiveness: The power of the structures of the model to represent concepts and to be interpreted as such concepts
 - For example, if the model can represent generalization / specialisation, composition, aggregation etc.
 - Behavioural expressiveness: Ability to represent the behavior of the concepts
 - Typically, if the model allows to express integrity constraints. For example, checks, assertions, etc.

Semantic Relativeness

- Definition: The system should support not just one conceptualization, but many. A conceptualization is never absolute, it is relative to the point of view of a user or group of people, because different persons perceive and conceive reality in differing ways. The system is not supposed to store a representation of each of these conceptualizations separately; it should store a single representation, encompassing all those conceptualizations, and avoiding redundancies
 - A powerful data model must not only be able to express the data source schemata and the global schema but also facilitate defining the mappings between them
 - A powerful algebra is needed to allow such transformations

Properties of a Canonical Data Model

- A canonical data model (CDM) must be:
 - More expressive than any of the data source data models
 - As other sources might be integrated in the future it must have a high degree of expressiveness
 - Must allow semantic enrichment
 - Express additional semantics to facilitate data integration (e.g., X same as Y)
 - Detect and solve semantic conflicts
 - Desirable characteristics:
 - The CDM must support *instances* and *classes*
 - Express rich relationships including generalisation / specialisation, aggregation, etc.
 - Must support arbitrary constraints
 - Must have a rich algebra allowing to derive new models
- Structural expressiveness**
- Behavioural expressiveness**
- Semantic Relativeness**

Additional Desirable Properties

- One basic structure: this way, there will be no modeling alternatives
 - If there are two or more (e.g., concepts and relationships) one may represent a real-world entity as a concept and another one as a relationship; which hinders integration due to the semantic heterogeneity
 - This might seem opposite to expressiveness, but this single structure may adopt different forms
- Multiple semantics: sometimes semantics cannot be expressed as generalisations / specialisations and alternative views must be presented