Schema Integration

Conceptual Database Design
Batini, Ceri, Navathe
Ch. 5

“A Comparative Analysis of Methodologies for Database Schema Integration”
Batini, Lenzerini, Navathe
ACM Computing Surveys, Vol 18, No 4, Dec 1986

Fundamentals of Database Systems
Elmasri/Navathe
sec. 16.2.2
Overview

• **Definition**
  – What is schema integration?
  – Merging two or more database schemas (models).

• **Problems**
  – Different methods of representing the same concepts, for example, naming concepts.

• **Strategies**
  – Schemas may be merged (i) in one go or (ii) in stages.

• **Process**
  – Three stages: (i) identify problems, (ii) resolve problems, (iii) merge schemas.

• **Resolving Conflicts**
  – Changing the names and the structure of entities.

• **Example**
  – See paper.
A *database schema* is the description of a database, for example, the entity-relationship model.

Batini et al define schema integration as “the process of merging several conceptual schemas into a *global conceptual schema* that represents all the requirements of the application”.

Schema integration is used to merge two or more database schemas into a single schema that can store data from both the original databases.

Schema integration is used when two or more existing databases must be combined, for example, when a new management information system is being developed.

Schema integration may be used when the process of database design is too large to be carried out by one individual. Two or more designers will build models of different parts of the database and use schema integration to merge the resulting models.

There are two major types of schema integration:

- **View Integration** View integration takes place during the design of a new database when user requirements may be different for each user group. View integration is used to merge different viewpoints into a single data model.

- **Database Integration** Database integration is used when two or more databases must be combined to produce a single schema, called a *global schema*.

*Ref: Batini, p119.*
Overview

- Definition
  - What is schema integration?
  - Merging two or more database schemas (models).

- Problems
  - Different methods of representing the same concepts, for example, naming concepts.

- Strategies
  - Schemas may be merged (i) in one go or (ii) in stages.

- Process
  - Three stages: (i) identify problems, (ii) resolve problems, (iii) merge schemas.

- Resolving Conflicts
  - Changing the names and the structure of entities.

- Example
  - See paper.
When two database schemas are designed by different designers using different user requirements, the resulting schemas will often present contrasting views of the same data.

In the example above, the relationship between employee and project in one database is represented as a relationship between employee, department and project in another database.

This situation might occur in an organisation that allows different departments to have different rules as to how employees are allocated to projects. For example, in one department employees may be assigned to projects while in another department employees may not be considered to be directly related to a project.

Ref: Baitin, sec. 5.1.
• Different databases may treat the same concepts in different ways.
• In the above example, the *publisher* concept is an entity in database 1 but an attribute in database 2.
• There are two situations that must be dealt with during schema integration:
  1. *When different concepts are modelled in the same way.* For example, in a university database staff and students may be represented by the entity *person* even though they are different concepts.
  2. *When the same concepts are modelled in different way.* For instance, the above example models the concept of a *publisher* as an entity and as an attribute.

*Ref: Batinì, sec. 5.1.*
Two database designs may be incompatible because mistakes were made in the initial design or there are different constraints placed on the data.

For instance, in the above example, the relationship between employee and project is represented as a one-to-many relationship in database 1 and as a many-to-many relationship in database 2.

This problem may be caused by mistakes made during the initial database analysis task or because users of the system have different working practices. For example, one department in an organisation, which works on small projects, may allocate one employee to a project but a different department, which works on large projects, may allocate many employees to a project.

During schema integration these different viewpoints must be reconciled.

Ref: Batini, sec. 5.1.
Overview

- Definition
  - What is schema integration?
  - Merging two or more database schemas (models).

- Problems
  - Different methods of representing the same concepts, for example, naming concepts.

- Strategies
  - Schemas may be merged (i) in one go or (ii) in stages.

- Process
  - Three stages: (i) identify problems, (ii) resolve problems, (iii) merge schemas.

- Resolving Conflicts
  - Changing the names and the structure of entities.

- Example
  - See paper.
The first strategy for integrating a set of schemas is to merge them all into a single large schema (called the *global schema*).

This approach would be difficult when the schemas are large or when there are a large number of schemas.

*Ref: Batini, sec. 5.2.*
The second strategy for schema integration is to integrate some of the schemas (e.g. two) and then to integrate the resulting schemas. This approach would be more appropriate when the schemas are complex or when there are a large number of schemas.

*Ref: Batini, sec. 5.2.*
Overview

• Definition
  – What is schema integration?
  – Merging two or more database schemas (models).

• Problems
  – Different methods of representing the same concepts, for example, naming concepts.

• Strategies
  – Schemas may be merged (i) in one go or (ii) in stages.

• Process
  – Three stages: (i) identify problems, (ii) resolve problems, (iii) merge schemas.

• Resolving Conflicts
  – Changing the names and the structure of entities.

• Example
  – See paper.
• The schema integration process starts with two or more schemas and involves three main stages:

1. **Conflict Analysis** During conflict analysis differences in the schemas are identified, for example, similar concepts that are represented in different ways.

2. **Conflict Resolution** During conflict resolution the conflicts identified during conflict analysis are resolved. For example, a common method of representing equivalent concepts will be decided upon. This process may involve discussing the problems with the users or correcting errors in the schemas.

3. **Schema Merging** During schema merging the schemas are merged into a single schema using the decisions made during the conflict resolution.

*Ref: Baitini, sec. 5.2.*
Overview

- **Definition**
  - What is schema integration?
  - Merging two or more database schemas (models).

- **Problems**
  - Different methods of representing the same concepts, for example, naming concepts.

- **Strategies**
  - Schemas may be merged (i) in one go or (ii) in stages.

- **Process**
  - Three stages: (i) identify problems, (ii) resolve problems, (iii) merge schemas.

- **Resolving Conflicts**
  - Changing the names and the structure of entities.

- **Example**
  - See paper.
Conflicts - Names

- **Synonyms**
  - Objects that are the same but have different names.
    - For example, *passenger* and *customer*.
- **Homonyms**
  - Objects that are different but have the same names.
    - For example, *publication (=book)* and *publication (=journal)*.

There are two types of name conflict that occur in a database schema:

- **Synonyms** When two similar concepts occur with different names. For example, two public transport databases may have entities called *passenger* and *customer*. These entities may be the same entity.
- **Homonyms** When two different concepts occur with the same name. For example, two publishing databases may have entities called *publication* but in one database a publication may be a book while in the other database a publication may be a journal.

Name conflicts cause a problem because information may be duplicated in the integrated database. It is important to identify those data items in each schema that actually represents the same concept or that should be represented using different structures in the integrated schema.

- Synonyms may be removed from the database by renaming the concepts so that they have the same name.
- Homonyms may be removed from the database by renaming the concepts so that they have different names.
- It may be possible to use a *superclass/subclass* relationship to avoid synonyms or homonyms.

*Ref: Batiní, sec. 5.3.1.*
Conflicts - Structural

- Identical concepts
  - Merged
- Compatible concepts
  - Representations are adapted and merged
- Incompatible concepts
  - Different cardinalities
  - Different identifiers
  - Reverse subset relationships

Structural conflicts occur when the actual method of representing the same concept in different databases is different or incompatible.

There are three cases:

- **Identical Concepts** When the same concept in different databases is represented in the same way they may be merged. For example, when an entity `publication` has the same structure and means the same in two databases the entities may be merged.

- **Compatible Concepts** When the same concept in different databases is represented in compatible ways they may be merged. For example, when an entity `publication` is represented by an attribute in one database and an entity in another database they may be merged by converting the attribute into an entity.

- **Incompatible Concepts** When the same concept in different databases is represented using different structures then it may be difficult to merge them directly. For example:
  - Relationships may have different cardinalities (i.e. one-to-many and many-to-many).
  - Primary keys may be different.
  - Set relationships may be reversed (e.g. projects contain programmes and programmes contain projects).

Incompatible designs must be resolved by re-analysing the data and adapting one or more of the schemas or by constructing a new, common representation.

*Ref: Batini, sec. 5.3.2.*
Overview

- **Definition**
  - What is schema integration?
  - Merging two or more database schemas (models).
- **Problems**
  - Different methods of representing the same concepts, for example, naming concepts.
- **Strategies**
  - Schemas may be merged (i) in one go or (ii) in stages.
- **Process**
  - Three stages: (i) identify problems, (ii) resolve problems, (iii) merge schemas.
- **Resolving Conflicts**
  - Changing the names and the structure of entities.
- **Example**
  - See paper.
Schema Integration

Example adapted from
“A Comparative Analysis of Methodologies for
Database Schema Integration”
Batini, Lenzerini, Navathe
ACM Computing Surveys, Vol 18, No 4, Dec 1986
Original Schemas
Step 1

Rename ‘Keywords’ to ‘Topics’
Step 2

Make the ‘publisher’ attribute an entity
Step 3

Two ‘Publisher’ entities merged together.

Two ‘Topic’ entities merged together.
Step 4

'Book' is a 'Publication'
Step 5

Remove relationships inherited from 'Publication'