Improving the Maintainability of Data Warehouse Designs: Modeling Relationships between Sources and User Concepts

Alejandro Maté
amate@dlsi.ua.es

Juan Trujillo
jtrujillo@dlsi.ua.es

Elisa de Gregorio
edg12@dlsi.ua.es

Il-Yeol Song
song@drexel.edu

International Workshop on Data Warehousing and OLAP
DOLAP'12
November 2nd, Maui, Hawaii

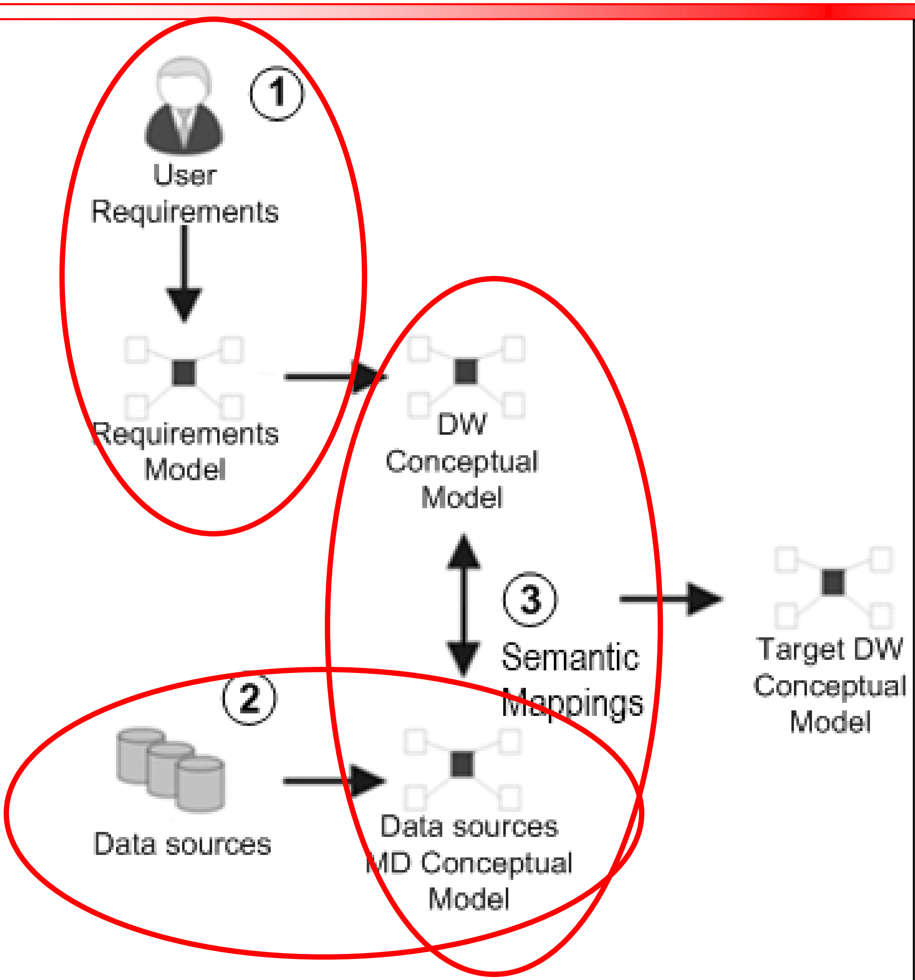
Universitat d'Alacant
Universidad de Alicante

Lucentia

Content

- Introduction
- Related Work
- Proposal
- Case study
- Conclusions & Future work

One Slide Summary



(1) Top-down, goal-oriented design
(2) Bottom-up, data-oriented design
(3) Capture semantic relationships:

- Attributes
- Hierarchy levels
- Dimensions

and derive Target DW model.:

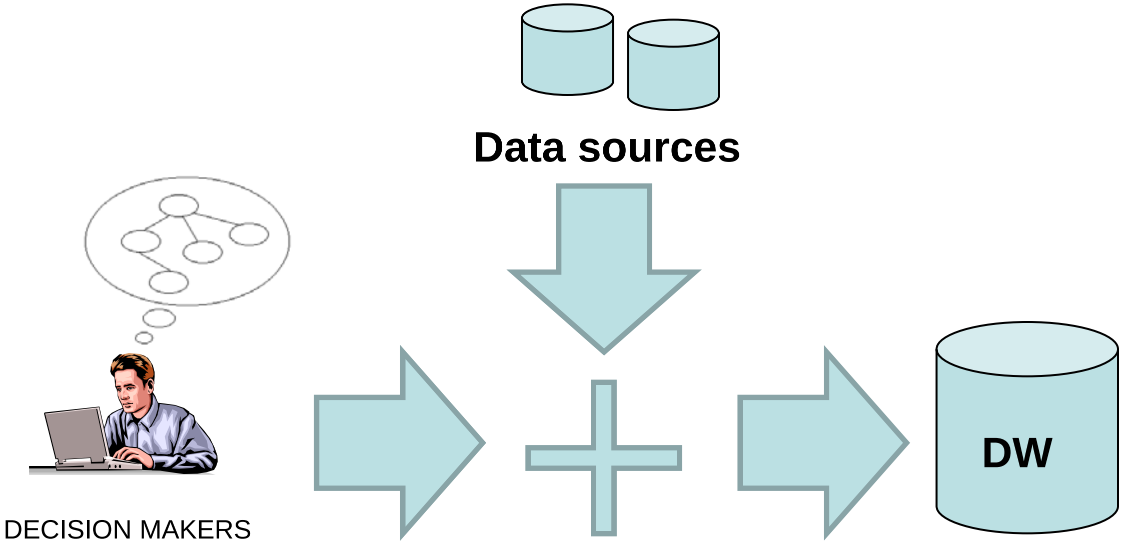
- Capture naming and structural mismatches
- Document the mappings
- Can evaluate the impact of changes, including which requirements may be affected
- Improve maintainability

Content

- Introduction
- Related Work
- Proposal
- Case study
- Conclusions & Future work

Introduction

- Developing a data warehouse requires information from users and data sources



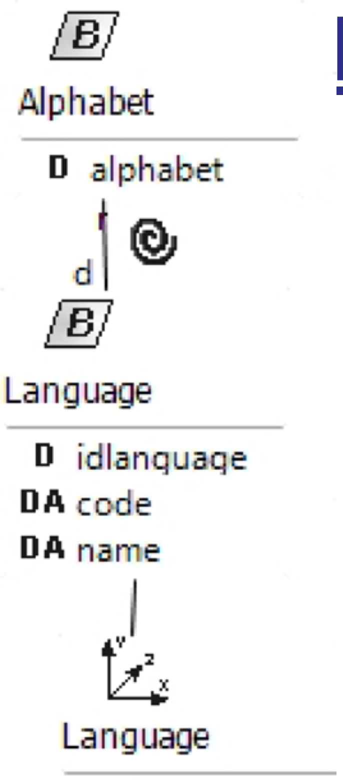
Introduction

- Motivation:
  - Hybrid DW development approaches **merge** user’s expectations with data source schemata

    [Mazón et al. 2009][Giorgini et al. 2008]
  - This task is **not trivial,** nor **well-documented**:
    - Naming conventions and structures usually **do not match**
    - May involve a **large number of tables**
    - Only documentation available are **ETL processes**
    - Considerations regarding **multidimensional aspects** are not recorded anywhere

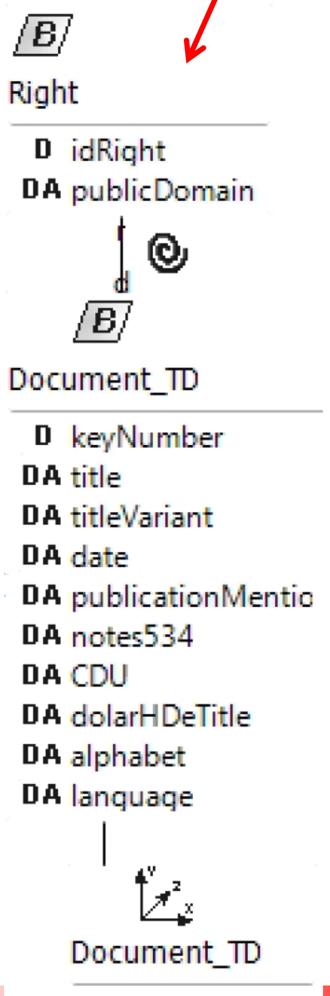
Introduction

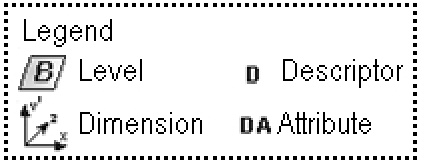
- What we **expect**:



What we **have**:

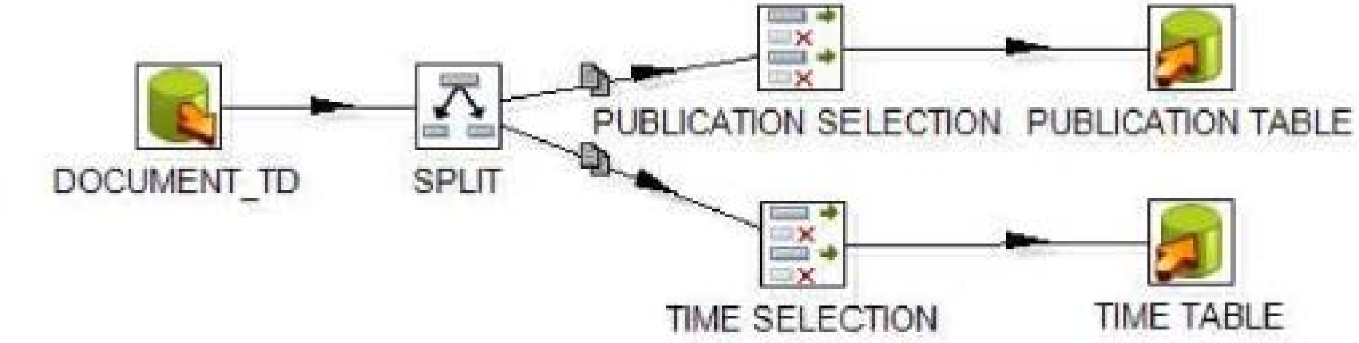
Data Sources





Introduction

- Information provided by ETL processes is **limited**:



Introduction

- Our long term goal:
  - Provide complete **traceability** of **every element** involved in the DW design process
- Objectives of this work:
  - Guide the DW designer on **identifying** the **relationships** in the reconciliation process
  - Provide a **formal framework** to identify these relationships
  - Allow DW designers to **accurately document** the reconciliation process

Content

- Introduction
- Related Work
- Proposal
- Case study
- Conclusions & Future work

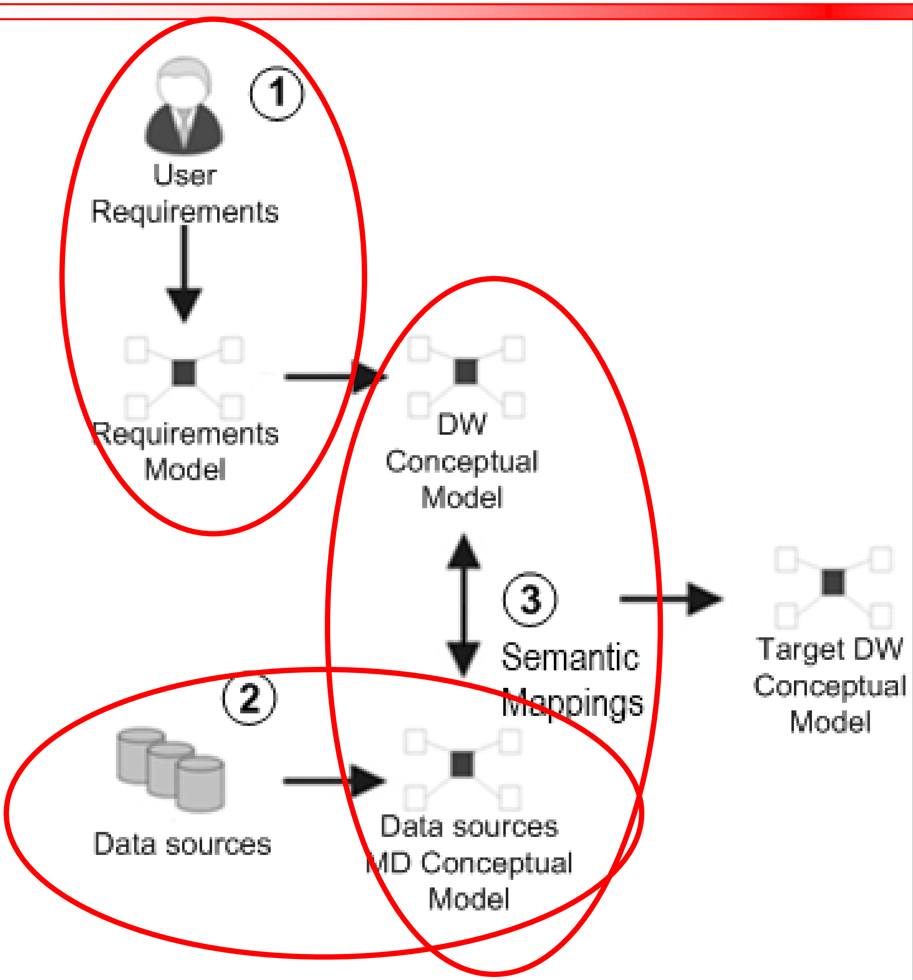
Related Work

- A **matching step** has been included in different hybrid methodologies *[Bonifati et al. 2001][Giorgini et al. 2008][Mazón et al. 2009]*
- This step **expects** that **naming conventions** are **maintained** from requirements to data sources
  - However, this is **rarely** the case *[Eckerson 2010]*
- Some proposals define a **common language (e.g., ontology)** to avoid this pitfall *[Bonifati et al. 2001][Romero et al. 2010]*
  - But there are also **structural differences**!!
- If **none** of the above apply, then, methodologies provide **no tools** for the designer **to tackle the problem**
  - The designer has to **redesign** the schema based on his experience

Content

- Introduction
- Related Work
- Proposal
- Case study
- Conclusions & Future work

Proposal



Modeling relationships between **expectations** and **data**

.

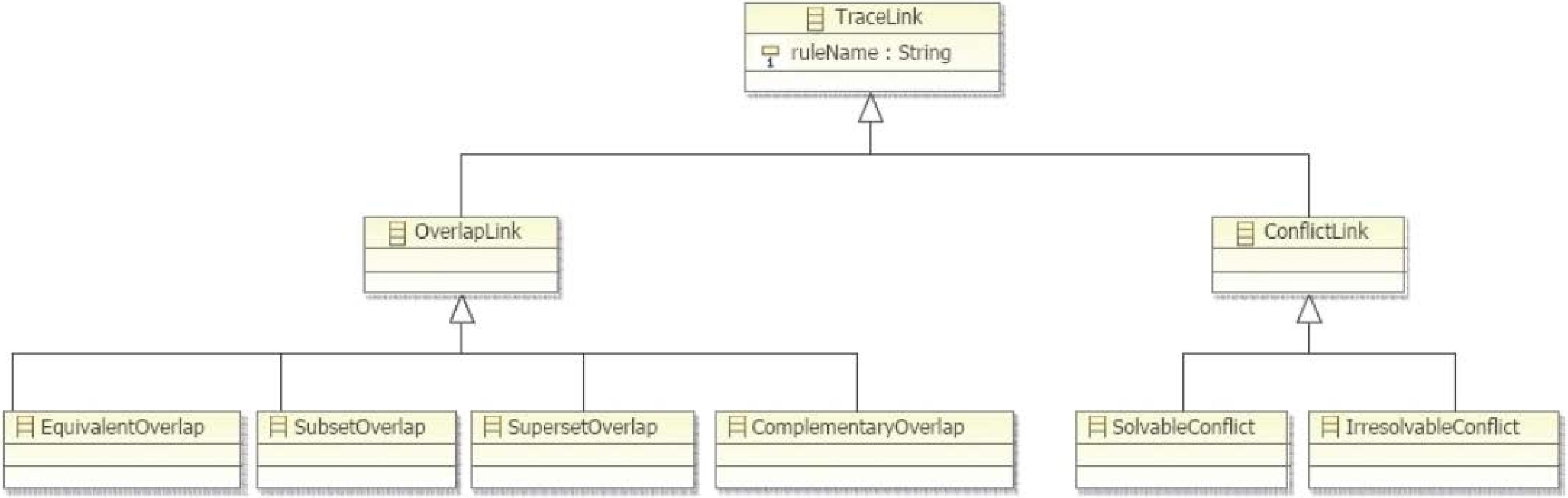
- Capture naming and structural mismatches
- Document the mappings
- Can evaluate the impact of changes, including which requirements may be affected
- Improve maintainability

Proposal

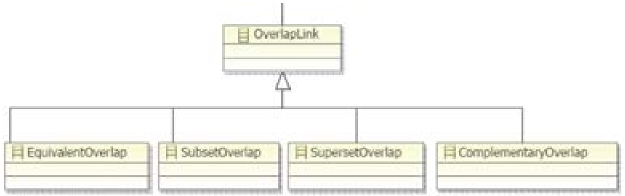
- Relationships are modeled at **three** different **levels**:
  - Attributes
  - Hierarchy Levels
  - Dimensions
- Using two basic concepts:
  - Overlap: No transformation needed
  - Conflict: A transformation must be found to provide adequate data

Proposal

- Specialized into six categories

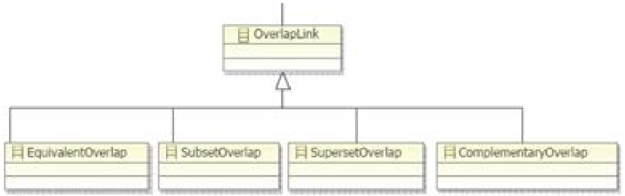


Proposal



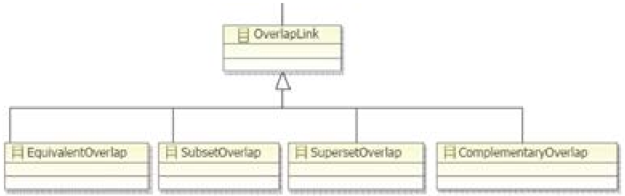
- Categories:
  - Categories describe the **semantics** of the relationships
  - Equivalent Overlap (EO): data available exactly matches our expectations, even if names are different
    - We expect a *Book* to have a *Title* and *Edition* number and we have a *Document* which has a *Title* and *EditionNumber*
  - Subset Overlap (UO): In one model, certain data is missing
    - *Document* only has a *Title* and does not have an *Edition* number

Proposal



- Categories:
  - Superset Overlap (SO): In one model, there are additional data on top of what we expected
    - *Document has Title, EditionNumber,* and *Language*
  - Complementary Overlap (CO): some expected information is missing while there is also additional data
    - We expect a *Book* to have a *Title* and *Edition* number, but the *Document* has a *Title* and *Language*
    - Structural differences usually cause multiple CO relationships appear

Proposal



- Categories:
  - Solvable Conflict (SC): the expected data is not available in the data sources but can be transformed
    - We included *Language* in our expectations, but we expected to retrieve a name, i.e. “Old English”. Instead, the data source actually provides a Language code “ang”. Using a code list we **can** obtain the name from the code
  - Irresolvable Conflict (IC): the conflict cannot be solved
    - If the code list was not available the previous transformation would not be possible

Proposal

- Attribute level:
  - Describe how much **information** is **provided**
  - Identify **missing attributes** and **transformations** required
  - Important for attributes used as **descriptors**

Proposal

- Examples:
  - Equivalent Overlap:
    - *keyNumber* includes the expected *idDocument* (EO). It stores ids by using a code for every document in the library.
  - Subset Overlap:
    - If *keyNumber* was missing information about certain documents.
  - Superset Overlap:
    - If*keyNumber* included information from documents in other libraries.

Proposal

- Examples:
  - Complementary Overlap:
    - If we expected *type* to include “handwritten” or “digital”. Instead, we have “handwritten”, “music composition”, “theater”.
  - Solvable Conflict:
    - *publicationMentio* stores information about the *place*, the *province,* and the *year* when a document was published, all mixed. It can be parsed (SC).
  - Irresolvable Conflict:
    - If *idDocument* expected titles as ids and, instead we had unrecognized codes stored in *keyNumber.*

Proposal

- Hierarchy Levels:
  - Level = (N,A),
    - A= a set of attributes and
    - N= semantic name of the level
  - Identify **concept mismatches in levels** could lead to different aggregated results!
  - Some aggregation levels may be **missing members** with no associated attributes
  - Some levels may not be transformable and thus require to be **substituted**

Proposal

- Examples:
  - Equivalent Overlap:
    - *Author* level: Both user expectations and data sources have the same set of attributes.
  - Subset Overlap:
    - A *Country* level; The data sources have only the *id* without the *name* of the *Country.*
  - Superset Overlap:
    - *Author* level: Data sources have bot only *Author* but also his/her *motherLanguage.*

Proposal

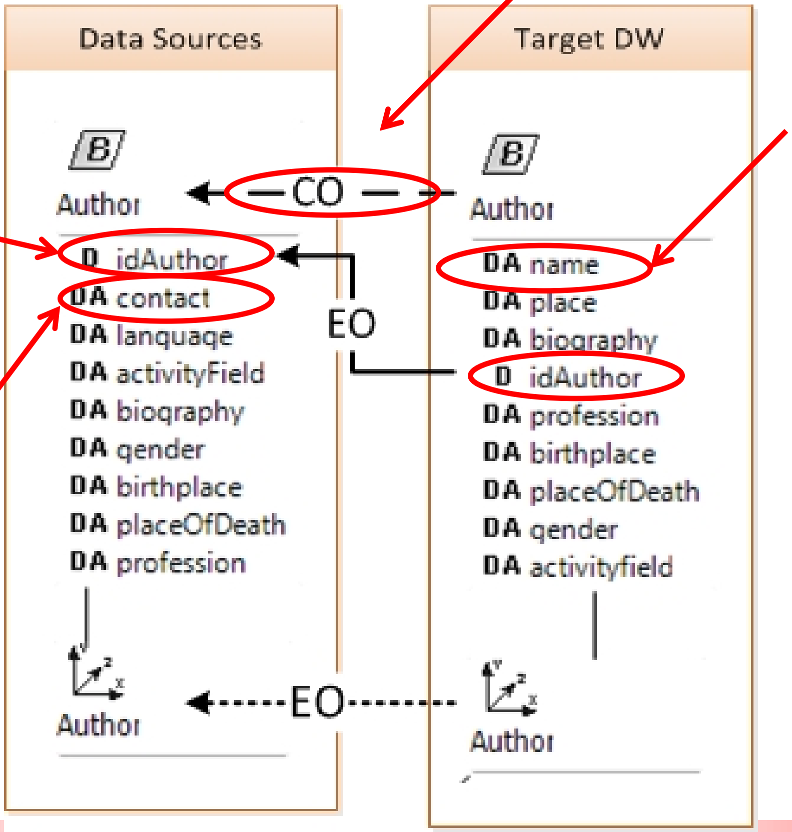
- Examples:
  - Complementary Overlap:
    - *Document Level: Document_TD in data sources* lacks a unique identifier, *uuid* , but includes information such as *notes534* and *date* instead
  - Solvable Conflict (level identification problem)
    - *Alphabet level in users:* Alphabet in *Document_TD is attribute.* Thus a transformation is needed
  - Irresolvable Conflict:
    - Language level: *Languages* in *Document_TD* has no id for the language and cannot be mapped.

Proposal

Example for Complementary Overlap between levels

- Graphical example:

1) The expected set of instances (ids) is provided



Legend

B Level

D Descriptor

Dimension

DA Attribute

3) Certain information is missing

2) Additional information is provided

Proposal

- Dimension level:
  - Identify **structural** differences between **dimensions hierarchies**
    - Can all the aggregation paths be created?
    - Is there any modification in the order of levels?
    - Is the granularity correctly defined?
  - Identify which dimensions are extracted from other dimensions

Proposal

- Examples:
  - Equivalent Overlap:
    - *Author* dimension: The data sources has the exact same levels we expected
  - Subset Overlap:
    - *User* dimension contains *User* and *User-Category* levels: but data sources has only *User* level
  - Superset Overlap:
    - *Publication* dimension: Data sources include an additional *State* level between *Provinces* and *Country levels*

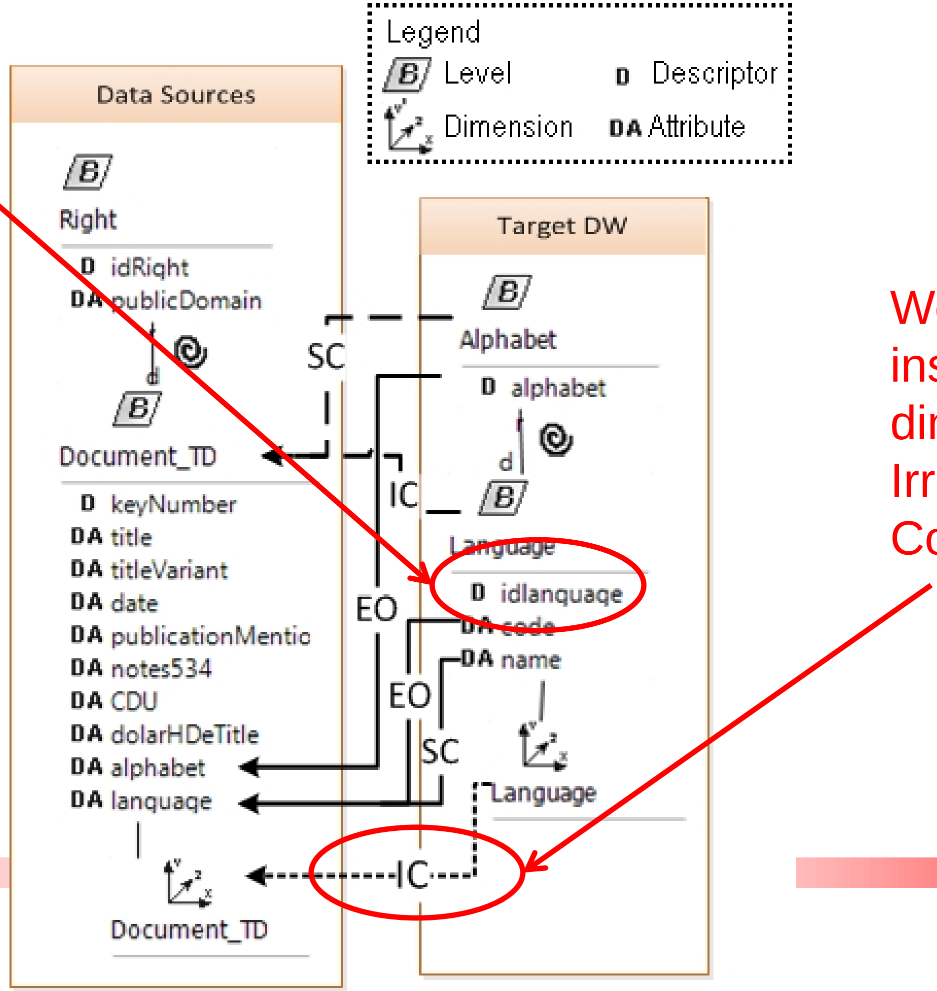
Proposal

- Examples:
  - Complementary Overlap:
    - *Document dimension has SupportForm* and *Type* levels*: Document_TD diemension in data sources lack them, but* includes the *Right* level.
  - Solvable Conflict:
    - *Document dimension* has *Format* as the second level.
    - *Format* dimension in data source has *Format* as its root
    - Hence, we have to apply a transformation to associate each format with its document
  - Irresolvable Conflict:
    - Language level: *Languages* in *Document_TD* has no id for the language and cannot be mapped.

Proposal

- Graphical example:

Descriptor (ID) for the lowest level is missing



We cannot obtain instances of the dimension → Irresolvable Conflcit (IC)

Content

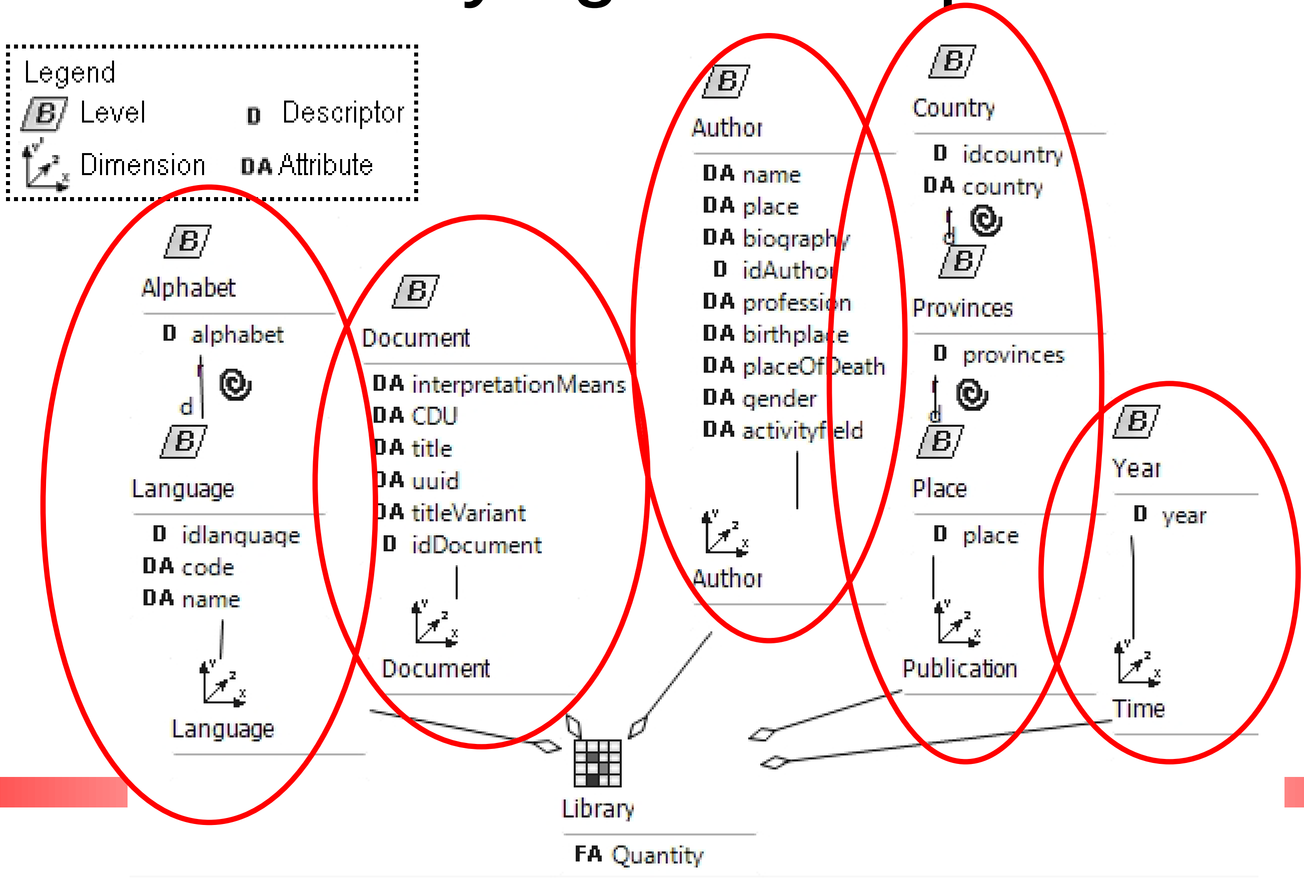
- Introduction
- Related Work
- Proposal
- Case study
- Conclusions & Future work

Case Study

- We applied our proposal to a real case study:
    - Integrating the information in the Digital Library at the University of Alicante
        - Combination of several data sources
        - Each data source is structured **according to a standard**
        - Necessity to **quickly** identify and **assess** how a **change** in the data sources affects the repository

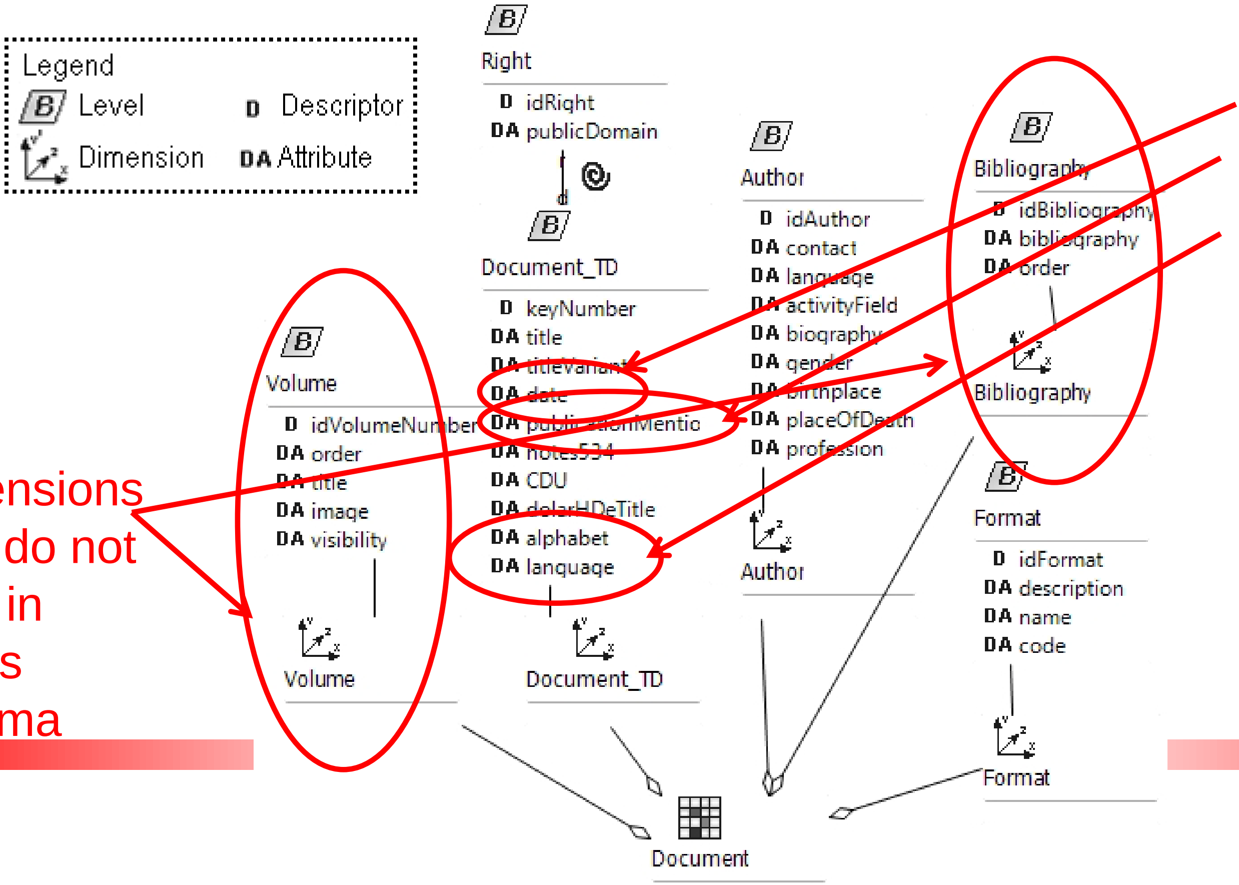
Case Study

- First step: obtain the multidimensional schema satisfying user requirements



Case Study

- Second step: obtain the multidimensional schema from Data source



Year?

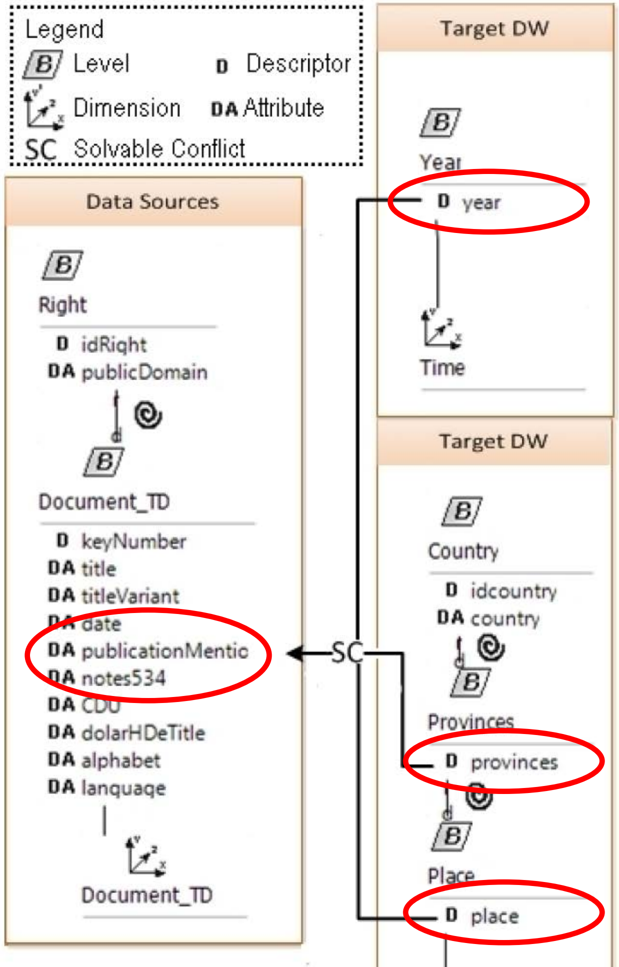
Publication?

Language?

New Dimensions That do not exist in user's schema

Case Study

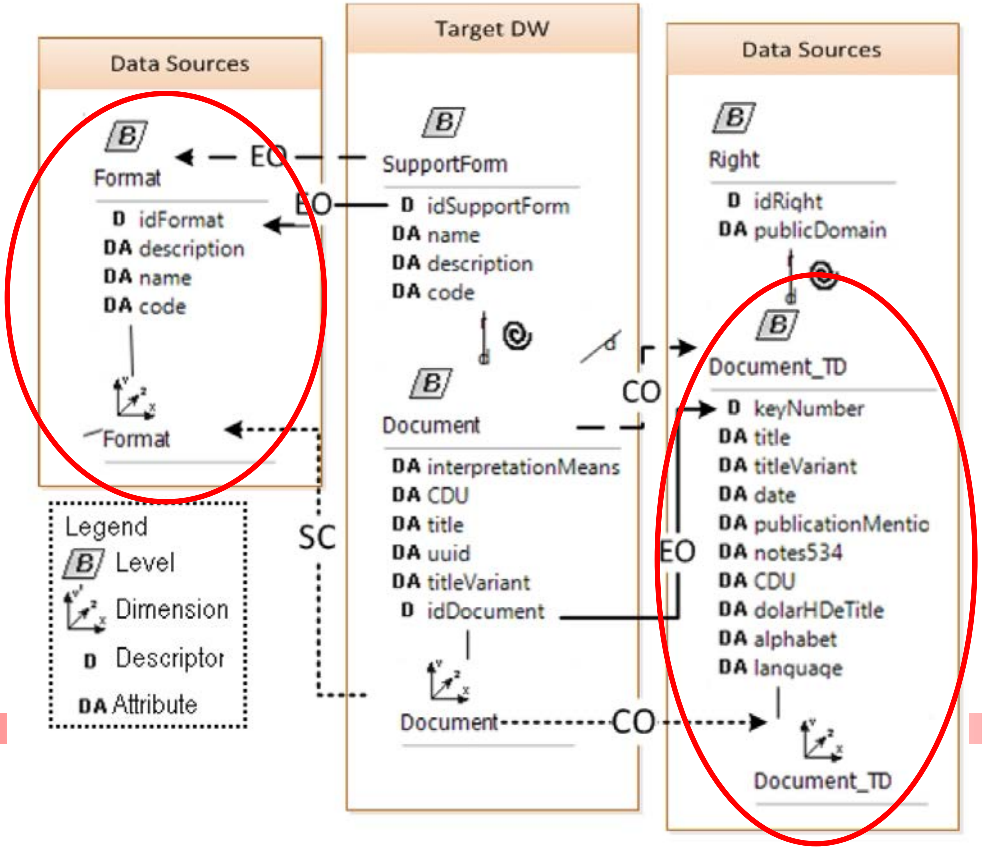
- Third step: relate elements by using our proposal



Case Study

- Third step: relate elements by using our proposal

Structural Differences: One Dimension is not enough!



Document information is obtained by combining Document_TD and Format dimensions

Case Study

- Final step: Analysis and Continuous Integration
  - All elements traced:
    - If a **new element** is **added**, we just **follow the previous steps** for its particular case
    - If an **element** is **removed** or **modified**, we **immediately know** which **elements** are **affected**
  - Mappings can provide us **additional** information:
    - We know which **elements** from the data sources are the ones **identifying** each **level** in the DW schema
    - We know which **requirements** are **only partially satisfied** as their concepts lack some information

Content

- Introduction
- Related Work
- Proposal
- Case study
- Conclusions & Future work

Conclusions & Future work

- Conclusions:
  - We have presented a formal framework to perform the reconciliation process
  - Our framework presents the following benefits:
    - **Explicit documentation** of the relationships between expectations and data sources not provided until now
    - As it is part of the DW traceability framework, it allows us to identify and **assess** the **impact** of any **change**
    - Allows us to incorporate new elements with a minimum impact on the DW schema

Conclusions & Future work

- Conclusions:
  - In addition, as a result of our approach, we can perform the following analysis:
    - Identify how many different sources are being employed for each requirement → Estimation of how much integration effort is required
    - As it is part of the DW traceability framework, we are able to identify which requirements can be really implemented and which ones cannot be (lack of data)
    - If new information is added, we can quickly identify if it makes viable those requirements which were previously unavailable
    - Provides important information for the decision maker, such as if certain information is missing (Subset Overlap), explaining why certain indicators are so low

Conclusions & Future work

- Future work:
  - Provide improved tool support for the approach
  - Define a series of metrics to evaluate the quality of the resulting DW and the impact of a change

Improving the Maintainability of Data Warehouse Designs: Modeling Relationships between Sources and User Concepts

Questions?

Alejandro Maté
amate@dlsi.ua.es

Juan Trujillo
jtrujillo@dlsi.ua.es

Elisa de Gregorio
edg12@dlsi.ua.es

Il-Yeol Song
song@drexel.edu

International Workshop on Data Warehousing and OLAP
DOLAP'12
November 2nd, Maui, Hawaii

Universitat d'Alacant
Universidad de Alicante

Lucentia