# Analyzing Dimension Mappings and Properties in Data Warehouse Integration

Domenico Beneventano, Marius Octavian Olaru, and Maurizio Vincini

Department of Engineering "Enzo Ferrari"
University of Modena and Reggio Emilia, Italy
firstname.lastname@unimore.it

**Abstract.** Data quality is one of the main issues both when building a Data Warehouse (DW) as when integrating two or more heterogeneous DWs. In the current paper, we perform extensive analysis of a mapping-based DW integration methodology and of its properties. In particular, the method is *coherent*, meanwhile under certain hypothesis it is able to preserve *soundness* and *consistency*. Furthermore, the necessary conditions for *homogeneity* will be discussed.

## 1 Introduction

Data Warehouse integration is the process of combining *multidimensional* information contained in two or more heterogeneous DWs by correctly identifying and integrating similar facts, measures and dimension categories. The process is required when two or more collaborating organizations need to be able to jointly take decisions based on strategic information obtained from all the participants of the collaboration. In such cases, however, *data quality* plays a crucial role, as decisions taken on incorrect information may be useless or may have a potentially negative impact on the organization. That is why any DW integration methodology must ensure basic quality properties, like coherency, soundness and consistency.

The problems when integrating two or more DWs are numerous. The information may be represented differently (e.g., a fact in one DW may be a dimension category in the other DW), with different values (e.g., a *month* may be represented by its name or by a number) or simply the same information is structured differently (e.g., different dimension categories).

In our previous work [1,2] we have proposed a mapping-based integration methodology that is able to generate mappings between dimension categories and to allow category and members import from similar dimensions. In the current paper we extend the previous work by formally reasoning about the properties of the mapping discovery and integration methodology. Interestingly, the mapping discovery step of the methodology is able to generate only mappings that are *coherent*, meanwhile the hypotheses under which *soundness* and *consistency* are preserved will be discussed.

The paper is structured as follows: Sect. 2 provides an overview of related work; Sect. 3 provides the preliminary discussion of dimension mappings and

R. Meersman et al. (Eds.): OTM 2013, LNCS 8185, pp. 616–623, 2013.

the properties that a mapping can have, meanwhile Sect. 4 provides the analytic discussion of the properties that are guaranteed and/or maintained while performing mapping discovery and dimension integration. Finally, Sect. 5 presents the conclusions of the current work.

## 2 Related Work

Data Warehouse integration has received little attention until recent years. Although most researchers proposed design methodology to facilitate the exchange of multidimensional information, recently a number of approaches tackled the DW integration problem in a systematic way.

For example, [3] proposes a mapping among dimensions defined as a partial function on the sets of dimension categories of the two mapped DWs. A mapping may have three properties, *coherency*, *soundness* and *consistency* that guarantee the correctness of the mapping among dimensions. The paper also introduces the *dimension algebra* as a way of manipulating DW dimensions, and two approaches to DW integration: a loosely coupled architecture and a tightly coupled architecture.

In [4], the authors propose a mapping methodology for DW elements (dimension levels and facts) by using *class similarity*. The approach presents some similarities with the integration methodology we previously developed [1,2]; however, meanwhile the authors in [4] rely on *class similarity* for the mapping process, we on the other hand use semantic similarity only as a validation step. Moreover, although the authors do not explicitly define and manage mapping properties, the methodology discard mappings that are not *coherent*.

In [5] the authors propose a semi-automatic definition of inter-attribute semantic mappings and transformation functions that can be used when performing low-level integration of different DWs.

## 3 Preliminaries

Although many models for representing DW dimensions have been proposed, for the purposes of the current paper the model presented in [6] will be used.

A *dimension schema* is a *directed acyclical graph* (dag) $H = (C, \nearrow)$, where $C$ is a finite set of *categories* and $\nearrow$ is a partial order relation over the set $C$. The partial order relation $\nearrow$ expresses the conceptual *roll-up* relations among categories. A *bottom category* $c_{bottom}$ is a category reachable from no other category of the schema: $\nexists c \in C$ such that $c \nearrow c_{bottom}$. A *hierarchy domain* is a dag $h = (M, <)$, where $M$ is the set of *members* of the hierarchy and $<$ a partial order relations over the set $M$. The reflexive and transitive closures of $\nearrow$ and $<$ will be denoted by $\nearrow^*$ and $<^*$ respectively. For example, in Fig. 1 *month* is a category, meanwhile "*Jan 2011*", "*Feb 2011*" are members of that category. Also, *month* $\nearrow$ *season* and "Jan 2011" $<$ "spring 2011", etc.

A *dimension* $d$ over a schema $(C, \nearrow)$ is a graph morphism $d : (M, <) \longrightarrow (C, \nearrow)$ such that $\forall x <^* y$ and $x <^* z$ such that $y \neq z$, then $d(y) \neq d(z)$.

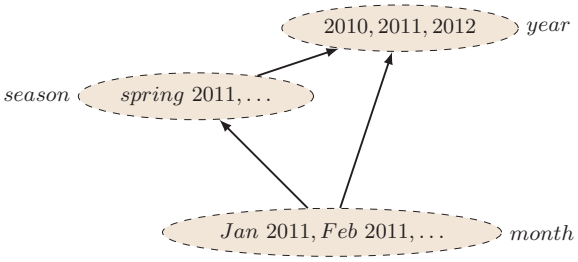


**Fig. 1.** A *time* dimension

Moreover, let $\mathfrak{m} : C \rightarrow \mathcal{P}(M)$ be a function that assigns each category the set of members mapped to the category by the graph morphism; in other words $\mathfrak{m}(c) = \{m \in M | d(m) = c\}$ is the set of members of $c$.

### 3.1 Dimension Mappings and Properties

A dimension mapping is a function that maps categories of one dimension to categories of another dimension [3]. Given two dimensions $d_1 : (M_1, <_1) \rightarrow (C_1, \nearrow_1)$ and $d_2 : (M_2, <_2) \rightarrow (C_2, \nearrow_2)$, a mapping is a *partial* function $\mu : C_1 \rightarrow C_2$.

The work in [3] defines properties that a mapping may have: *coherency*, *soundness* and *consistency*.

**Coherency** guarantees that mapped dimension categories roll-up to similar categories. Formally, $\mu$ is coherent if $\forall c_i, c_j \in C$, $c_i \nearrow_1^* c_j \Leftrightarrow \mu(c_i) \nearrow_2^* \mu(c_j)$.

**Soundness** is guaranteed whenever mapped categories contain the same members, meaning that $\mathfrak{m}(c) = \mathfrak{m}(\mu(c))$ for all $c \in C_1$.

**Consistency** states that members of mapped categories roll-up to members of categories that are also mapped by $\mu$. In other words, $\forall m_{i_1}, m_{j_1} \in M_1$ such that $m_{i_1} <_1^* m_{j_1}$ then $\forall m_{i_2} \in \mathfrak{m}(\mu(d(m_{i_1})))$ such that $m_{i_1} = m_{i_2}$, then $\exists m_{j_2} \in \mathfrak{m}(\mu(d(m_{j_1})))$ such that $m_{i_2} <_2^* m_{j_2}$.

Furthermore, a mapping that is coherent, sound and consistent is called a *perfect mapping*.

**Schema Homogeneity**
In [6] a schema is defined *homogeneous* if for all categories $c_i \nearrow c_j$, if a member of $c_i$ rolls-up to a member of $c_j$, than all members of $c_i$ roll-up to a member of $c_j$. Otherwise, it is *heterogeneous*.

### 3.2 A Dimension Integration Methodology

In our previous work [1,2] we have presented a mapping and integration methodology that is able to: (a) *map* dimension categories and (b) *import* dimension members from one dimension to another.

The mapping generation step uses the graph-like structure of dimension and cardinality-related properties to map categories. Subsequently, semantic similarity

is used to discard *unreliable* mappings by computing a similarity score with the *MELIS* [7] methodology.

The second step of the methodology imports categories and member from one dimension to another by using the following approach. Let $d_1$ and $d_2$, be two dimensions, and $c_i, c_j \in C_1$ such that $c_i \nearrow_1 c_j$ and $c'_i \in C_2$. Let $\xi$ be a mapping generated by the first step of the methodology. If $\xi(c_i) = c'_i$ and $\xi(c_j) \notin C_2$, then $d_2$ is augmented with the category $c_j$ and with the roll-up relations derived from the mappings. A new dimension $d_2^\sharp : (M_2^\sharp, <_2^\sharp) \rightarrow (C_2^\sharp, \nearrow_2^\sharp)$ is derived, where $C_2^\sharp = C_2 \cup \{c_j\}$; $\nearrow_2^\sharp$ is extended to include the relations between $c_j$ and the categories of $C_2$. Also, $M_2^\sharp = M_2 \cup \{\mathfrak{m}(c_j)\}$ (see Fig. 2) and relation $<_2$ is extended to $<_2^\sharp$ in order to include the relations between the newly imported members and the initial members of the categories in $d_2$. Formally, for every $c_i, c_j \in C_1$ and $c'_i, c'_j \in C_2^\sharp$ such that $c_i \nearrow_1 c_j$ and $c'_i \nearrow_2^\sharp c'_j$, if $\xi(c_i) = c'_i$, then for all $m_i \in \mathfrak{m}(c_i)$ and $m_j \in \mathfrak{m}(c_j)$ such that $m_i <_1 m_j$, and for all $m'_i \in \mathfrak{m}(c'_i)$ such that $m_i = m'_i$, then it must be that $m_j \in M_2^\sharp$ and $m'_i <_2^\sharp m_j$ (see Fig. 2). The mapping $\xi$ is extended to include the newly imported category. A new mapping $\xi^\sharp : C_1 \rightarrow C_2^\sharp$ is generated as follows:

$$\xi^\sharp(c) = \begin{cases} \xi(c), & \text{if } c \neq c_j \\ c_j, & \text{if } c = c_j \end{cases}$$

The members of $c'_j$ are imported using an approach based on the *RELEVANT* [8] clustering methodology. Note that, for the sake of simplicity, the example in Fig. 2 does not highlight the advantages of using *RELEVANT*, which was chosen for its ability to combine information from more than one dimension and to discriminate between incorrect members by using clustering techniques rather than direct equality of the members. In our methodology, *RELEVANT*

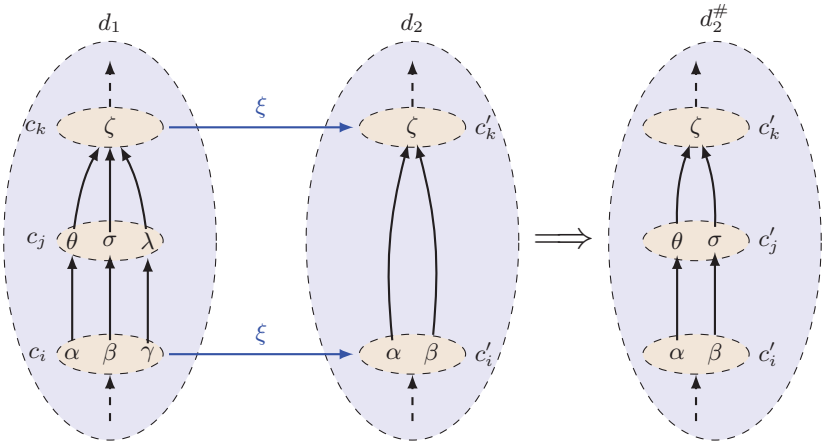


**Fig. 2.** The *category* and *member* importation rule

is used by enforcing that clustered members from mapped dimension categories roll-up to members of the same cluster (as computed by *RELEVANT*) of members of mapped categories.

## 4 Property Analysis

In this section we will analyze the properties that are guaranteed and preserved by the proposed dimension integration methodology. To this end, we will denote by $\xi$ a mapping generated by the first step of such methodology (see previous section). Also, the proofs of all Theorems and Corollaries are reported in [9].

### 4.1 Coherency, Soundness and Consistency Check

The first important result is that the mapping generated by the presented methodology is always *coherent*.

**Theorem 1.** *The mapping $\xi$ is coherent.*

The proof of the theorem is presented in [9] and it relies on graph-theory to demonstrate that the mappings are coherent by construction.

Although the first step of the integration methodology produces a coherent mapping, soundness and consistency are guaranteed only in certain cases. In order to verify whether soundness is verified, two steps must be performed. First, the initial mapping must be checked. Formally, for all categories $c \in C_1$, it must be that $\mathfrak{m}(c) = \mathfrak{m}(\xi(c))$. This is a simple inclusion test that will be analyzed no further. Secondly, the soundness and consistency property must be verified when performing the category and member importation.

The following theorem provides a sufficient condition for guaranteeing soundness and consistency when importing categories and members.

**Theorem 2.** *If $\xi$ is sound and consistent, then $\xi^\sharp$ is also sound and consistent.*

Theoreme 2 provides a *sufficient*, but not *necessary* condition for soundness and consistency. In fact, there may be cases when mapping two dimensions where the initial mapping $\xi$ is neither sound nor consistent, but the final mapping $\xi^\sharp$ becomes sound and/or consistent. For example, Fig. 3 provides two dimensions and a mapping $\xi$ that is neither sound nor consistent, as $\mathfrak{m}(c_j) \neq \mathfrak{m}(c'_j)$ (the member $\sigma$ belongs to $\mathfrak{m}(c_j)$ but not to $\mathfrak{m}(c'_j)$) and member $\beta$ rolls-up to member $\alpha$ in dimension $d_1$, but rolls-up to no member of $c'_j$ in dimension $d_2$. Note that dimension $d_2$ is also heterogeneous. Assuming the methodology generated the mapping $\xi$ (see Fig. 3), step 2 of the methodology renders the mapping both sound and consistent.

The reason for which soundness and consistency are considered together is that they are closely related. In some cases (not all) soundness follows from consistency. The following corollary states a relationship between consistency and soundness of a mapping relation.

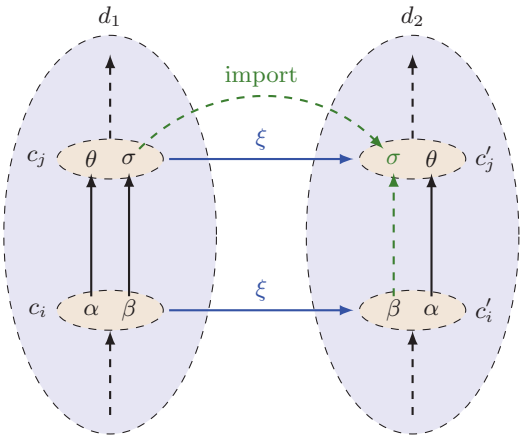


**Fig. 3.** No sound → sound mapping

**Corollary 1.** *If $\xi$ maps only pairs of categories $c_i$ and $c_j$ such that $c_i \nearrow c_j$ and $\xi$ is consistent, than $\xi$ is also sound.*

Moreover, a mapping $\xi$ yielding all three properties remains a *perfect* mapping after the integration methodology.

**Corollary 2.** *If $\xi$ is a perfect mapping, then $\xi^\sharp$ is also a perfect mapping.*

### 4.2 Checking Homogeneity

Unfortunately, homogeneity is not preserved when integrating different DW dimensions. Not even the case when integrating two homogeneous dimensions can ensure that the derived dimension is homogeneous.

Interestingly, heterogeneity is also not preserved. There may be the case when a homogeneous dimension is obtained when integrating two heterogeneous dimensions. For example, in Fig. 4 when integrating members from dimension $d_1$ to dimension $d_2$ (both heterogeneous), the newly derived dimension $d_2^\#$ is homogeneous. In this later case, the instance of dimension $d_2$ is completed with information from $d_1$. A situation like the one described in Fig. 4 may be encountered in real life cases when analysts decide to model the same information differently, or when information is partially missing by choice or by error. For example, categories $c_j$ and $c'_j$ may represent the region of a city (categories $c_i$ and $c'_i$), that was omitted for some cities in $d_1$ (member $\beta$) or for other members in $d_2$ (member $\alpha$). The missing information is thus derived from the other dimension.

In some circumstances, homogeneity may be maintained when integrating two different homogeneous dimensions. The following theorem provides a sufficient condition to guarantee the preservation of homogeneity.

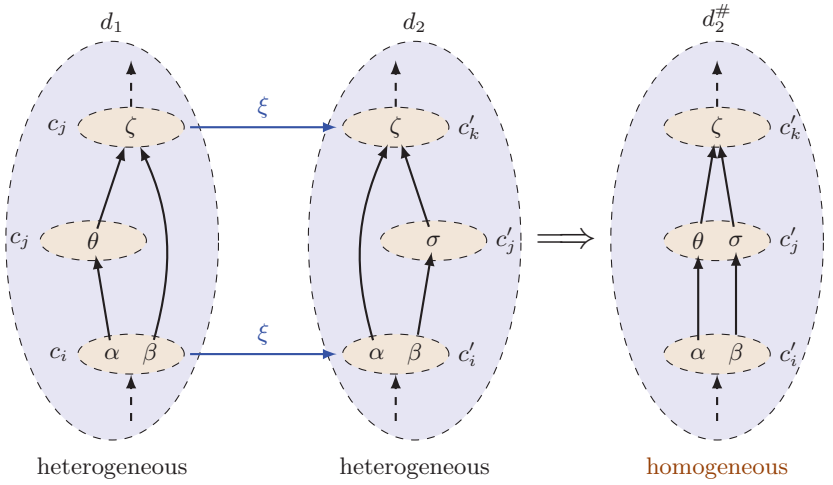


**Fig. 4.** Heterogeneous $\Longrightarrow$ Homogeneous

**Theorem 3.** *If $d_1$ and $d_2$ are homogeneous and $\mathfrak{m}(c'_i) \subseteq \mathfrak{m}(c_i)$, then $d_2^{\sharp}$ is also homogeneous.*

An interesting observation may be drawn from Theorem 3. It turns out that when integrating two homogeneous dimensions $d_1$ and $d_2$ with bottom categories $c_{\text{bottom}_1}$ and $c_{\text{bottom}_2}$, if $\mathfrak{m}(c_{\text{bottom}_1}) = \mathfrak{m}(c_{\text{bottom}_2})$, then by importing categories and members from one dimension to another, the newly obtained dimensions $d_1^{\#}$ and $d_2^{\#}$ are identical, a part from the names of the categories. In other words, there will be a total mapping $\varkappa : C_1^{\#} \to C_2^{\#}$ that is *perfect*. Furthermore, $\varkappa^{-1}$ is also a *perfect* mapping. The newly derived dimensions $d_1^{\#}$ and $d_2^{\#}$ are identical to the one derived in [3] using the *tightly coupled* approach.

**Corollary 3.** *If $\mathfrak{m}(c'_i) \not\subseteq \mathfrak{m}(c_i)$ and $c_j \notin C_2$, then $d_2^{\#}$ is heterogeneous.*

## 5 Conclusions

In the current paper, a formal analysis of a mapping based DW integration methodology and its properties have been performed. The presented methodology is able to generate mappings that are *coherent*, which in turn allow correct aggregation of information from the different DWs. Moreover, under specific constraints, after performing the integration steps, the mapping may also be rendered *sound* and *consistent*.

Coherency and consistency are properties related to roll-up relations, thus a mapping satisfying both properties ensures that the integrated information can be correctly aggregated (or disaggregated). This observation is relevant as the multidimensional data is usually explored along aggregation patterns, drilling-down or rolling-up from a starting analysis point.

On the other hand, soundness states that the mapped dimension categories contain the same members, which in turn give analysts the possibility of executing meaningful drill-across queries that would otherwise be impossible if the related categories contained distinct members.

Finally, although some researchers allow the design of heterogeneous dimensions, we on the other hand consider that homogeneity allows a more clear representation of multidimensional data, both for designers and analysts as for business people that may have a simpler perception of the underlying DW model. The current paper analyzed schema heterogeneity and provided a sufficient condition for maintaining homogeneity that is a necessary condition for summarizability[10] and for materializing views as a mean of optimizing response time when executing dependent queryes[11].

## References

1. Bergamaschi, S., Olaru, M.O., Sorrentino, S., Vincini, M.: Dimension matching in Peer-to-Peer Data Warehousing. In: 16th IFIP WG 8.3 International Conference on Decision Support Systems, Anávissos, Greece, pp. 149–160 (2012)
2. Guerra, F., Olaru, M.-O., Vincini, M.: Mapping and Integration of Dimensional Attributes Using Clustering Techniques. In: Huemer, C., Lops, P. (eds.) EC-Web 2012. LNBIP, vol. 123, pp. 38–49. Springer, Heidelberg (2012)
3. Torlone, R.: Two approaches to the integration of heterogeneous data warehouses. Distributed and Parallel Databases 23(1), 69–97 (2008)
4. Banek, M., Vrdoljak, B., Tjoa, A.M., Skocir, Z.: Automated Integration of Heterogeneous Data Warehouse Schemas. IJDWM 4(4), 1–21 (2008)
5. Bergamaschi, S., Guerra, F., Orsini, M., Sartori, C., Vincini, M.: A semantic approach to ETL technologies. Data & Knowledge Engineering 70(8), 717–731 (2011)
6. Hurtado, C.A., Gutierrez, C., Mendelzon, A.O.: Capturing summarizability with integrity constraints in OLAP. ACM Transactions on Database Systems 30(3), 854–886 (2005)
7. Bergamaschi, S., Bouquet, P., Giacomuzzi, D., Guerra, F., Po, L., Vincini, M.: An Incremental Method for the Lexical Annotation of Domain Ontologies. Int. J. Semantic Web Inf. Syst. 3(3), 57–80 (2007)
8. Bergamaschi, S., Sartori, C., Guerra, F., Orsini, M.: Extracting Relevant Attribute Values for Improved Search. IEEE Internet Computing 11(5), 26–35 (2007)
9. Beneventano, D., Olaru, M.O., Vincini, M.: Dimension Mapping and Properties (Technical Report) (2013), `http://www.dbgroup.unimo.it/files/ODBASE.pdf`
10. Lenz, H.J., Shoshani, A.: Summarizability in OLAP and Statistical Data Bases. In: SSDBM, pp. 132–143 (1997)
11. Harinarayan, V., Rajaraman, A., Ullman, J.D.: Implementing data cubes efficiently. In: Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data - SIGMOD 1996, pp. 205–216. ACM Press, New York (1996)