A Framework for the Identification of Integrity Constraints in Spatial Datacubes

Mehrdad Salehi, Yvan Bédard, Mir Abolfazl Mostafavi, Jean Brodeur

Center for Research in Geomatics, Department of Geomatics, Laval University

Research Context

• Spatial data warehouses are an integral part of modern decision support systems

• These data warehouses are often structured based on datacubes

 + = Spatial datacube

• Data quality in spatial datacubes is important since they are used as the database backend in large organization

• Integrity constraints (ICs) play an important role in improving one of the elements of data quality, *i.e.*, logical consistency (ISO/TC211 - 19113)

• ICs are the rules that are imposed on data in order to prevent the appearance of inconsistence data in databases

Research Problem

There is no model for spatial datacubes that explicitly includes ICs:

Spatial datacube designers identify ICs based on non-systematic and *ad-hoc* ways

Research Objective

To provide a framework for spatial datacube designers so that they easily identify the ICs of spatial datacubes

Classification of the ICs in Spatial Datacubes Based on Their Nature
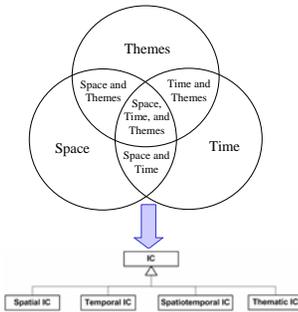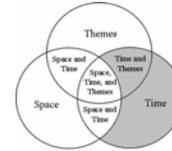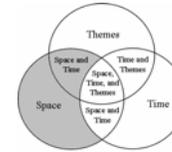

1. Spatial IC: An IC whose assertion includes spatial concepts and maybe thematic concepts

Example: The distance between a school and a gas station must be greater than 300 meters.

2. Temporal IC: An IC whose assertion includes temporal concepts and maybe thematic concepts

*Example*: The construction date of a building must be before its renovation date.

3. Spatiotemporal IC: An IC whose assertion includes spatiotemporal concepts or spatial and temporal concepts together and maybe thematic concepts

Example: If the acceleration of a moving-point is greater than zero, its speed must increase.

**4. Thematic IC**: An IC whose assertion includes thematic concepts

Example: The number of floors of a building must be greater than zero.
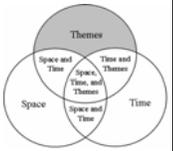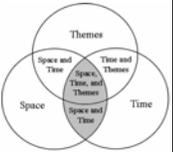
Identification of the ICs in Spatial Datacubes

A formal model for spatial datacubes

Component	Definition	Formalization
Attribute	Describes a level	a_i
Level	Analysis granularity along a dimension	$l = \{a_1, \ldots, a_n\}$
Dimension	Axis of analysis	$d = (\{l_1, \ldots, l_n\}, \prec)$
Member	Instance of a level	$m = (\{a_1, \ldots, a_n\}, \{v_1, \ldots, v_n\}, \ldots)$
Measure	Attribute analyzed taking into account dimension levels	ms_i
Hyper-cell	A model for several facts	$hc = (L, MS)$
Datacube Schema	Describes the structure of a datacube	(D, MS, HC)
Fact	Describes an event for decision making	$f = (M, V)$
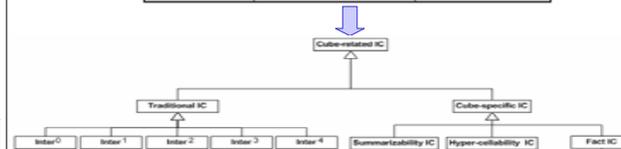
**1. Traditional ICs**: Prevent the inconsistencies that can happen when integrating members

Example: If two members of the level "center" are of the type "airport" and they exist during the same period of time, then the distance between these two members must be at least 40 km.

2. Summarizability ICs: Prevent meaningless data aggregation

Example: The measure "postal code" in not summarizable.

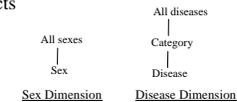**3. Hyper-cellability ICs**: Prevent the creation of meaningless hyper-cells

Example:
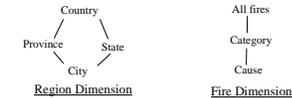
Hyper-cellability IC: There is no hyper-cell that includes two levels "Province" and "Cause" together.

4. Fact ICs: Prevent undesirable facts

Example:

Fact IC: There is no fact that includes the disease "Prostate Cancer" and the sex "Female".

Conclusion

• ICs play an important role in improving data quality

• Due to the absence of a framework for identifying the ICs of spatial datacubes, these ICs had been treated based on non-systematic ways

• We identified the necessary ICs for spatial datacubes

• Based on this solution, spatial datacubes designers can identify the ICs of spatial datacubes *a priori* and systematically