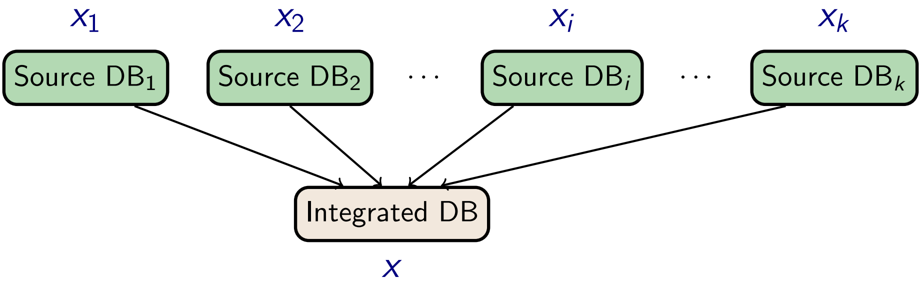
Integration Integrity for Multigranular Data

Stephen J. Hegner
Umeå University, Sweden

M. Andrea Rodríguez
University of Concepción, Chile

ADBIS 2016
20th East-European Conference on
Advances in Databases and Information Systems
Prague, Czech Republic
30 August 2016



Task: Several source DBs are to be combined into a single integrated DB.

- Assume that each source DB is *locally consistent*.

Consistency problem: There may exist additional *global constraints* which apply when all source DBs are considered together.

Example constraint: $\sum_{i=1}^{k} x_i = x$.

- This constraint arises only in a context in which all data items in $\{x_1, x_2, \ldots, x_k, x\}$ occur.
  - In other words, only on the integrated DB.

Consistency of Multigranular Data

Source database 1		
Place	Time	Births
Reg_I	*Q1Y2014*	n_1
Reg_II	*Q1Y2014*	n_2
...	...	...
Reg_XV	*Q1Y2014*	n_{15}

Source database 2		
Place	Time	Births
Chile	*Q1Y2014*	b_1
Chile	*Q2Y2014*	b_2
Chile	*Q3Y2014*	b_3
Chile	*Q4Y2014*	null

Source database 3		
Place	Time	Births
Chile	*Y2012*	b_{12}
Chile	*Y2013*	b_{13}
Chile	*Y2014*	b_{14}
Chile	*Y2015*	b_{15}

Disjointness constraints are central to this work:

- Chile is the disjoint union of its fifteen regions:

$$\bigsqcup^{\perp}_{\text{Place}} \{Reg\_R \mid I \leq R \leq XV\} = Chile$$

- Year 2014 is the disjoint union of its quarters:

$$\bigsqcup^{\perp}_{\text{Time}} \{QxY2014 \mid 1 \leq x \leq 4\} = Y2014$$

Consequences:

- $\sum_{i=1}^{15} n_i = b_1$ (constraint for integration of DB 1 and DB 2).
- $\sum_{i=1}^{3} b_i \leq b_{14}$ (constraint for integration of DB 2 and DB 3).

- Even to integrate just DB 1 and DB3, need $\sum_{i=1}^{15} n_i \leq b_{14}$ to hold.

The Concept of a TMCD

Source database 1		
Place	Time	Births
Reg_I	*Q1Y2014*	n_1
Reg_II	*Q1Y2014*	n_2
...	...	...
Reg_XV	*Q1Y2014*	n_{15}

Source database 2		
Place	Time	Births
Chile	*Q1Y2014*	b_1
Chile	*Q2Y2014*	b_2
Chile	*Q3Y2014*	b_3
Chile	*Q4Y2014*	null

Source database 3		
Place	Time	Births
Chile	*Y2012*	b_{12}
Chile	*Y2013*	b_{13}
Chile	*Y2014*	b_{14}
Chile	*Y2015*	b_{15}

- For simplicity, the source databases are assumed to have the same relational structure, but at different granularities.

Thematic multigranular comparison dependencies: *TMCDs* generalize ordinary FDs for the multigranular framework.

- The notation for an example *TMCD* is shown below.

variable attribute | fixed attribute | op is equality | thematic attribute

$$\text{Place}\ \underline{\text{Time}} \xrightarrow[(\perp,1)]{\doteq} \langle \text{Births}: \langle \theta, \textstyle\sum, \tau \rangle \rangle$$

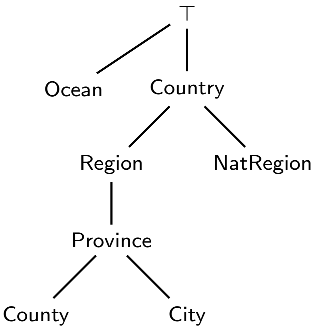
Form is $\sqcup_{\text{Place}} \langle \text{set} \rangle \circledast \langle \text{value} \rangle$

thematic order | aggregation operator | aggregation tolerance

Modelling Multigranular Data — Granularities

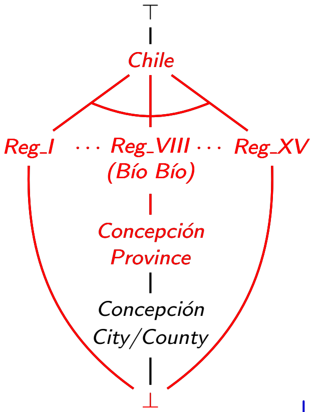
- In the classical relational model, the attribute domains are *flat*.
- In the multigranular model, the attribute domains have partial-order (*poset*) structure.
- *Granularities* are the *types*, while *granules* are the *domain values*.

Example *granularities* for the attribute Place:



- Going up results in *coarser* granularity.
- There is always a coarsest granularity $\top$.
- Every nonempty set of granularities has at least one minimal upper bound (MUB).
- No other algebraic structure (join, meet, complement, $\perp$) is utilized.
- An ordinary (flat) attribute is recaptured via just $\top$ plus the single, main granularity.

Modelling Multigranular Data — Granules



- Shown is a small fragment of the granule structure for attribute Place.
- The poset is *bounded*: $\bot$ and $\top$ are always present.
- There are three types of *rules*:

  Ordinary subsumption:

  $$\textit{Concepción Province} \sqsubseteq \textit{Reg\_VIII}$$

  Join: $\bigsqcup_{\mathsf{Place}}\{Reg\_R \mid I \leq R \leq XV\} = Chile$

  Binary disjunction: $Reg\_R \wedge Reg\_S = \bot$

$\stackrel{\perp}{\bigsqcup} = \bigsqcup$ + pairwise binary disjunction:

$$\stackrel{\perp}{\bigsqcup}_{\mathsf{Place}}\{Reg\_R \mid I \leq R \leq XV\} = Chile$$

- The structure must complete to a *distributive* lattice.
- This is always satisfied in practice for spatio-temporal attributes.
- Join corresponds to union and meet to intersection in that case.

The Interaction of Granularities and Granules

- For granular attribute A, granules are assigned to granularities via a *granulated domain assignment*.
- $\mathrm{GrtoDom}_A(G)$ = granules of granularity G.
  Example: $\mathrm{GrtoDom}_{\mathrm{Place}}(\mathit{Region}) = \{\mathit{Reg\_R} \mid I \leq R \leq XV\}$.
- Granularity $\top$ consists of granule $\top$.
- Every granule except $\perp$ belongs to at least one granularity.
  Example: *Concepción* City = *Concepción* County (same granule).
- The granules $\mathrm{GrtoDom}_A(G)$ of a given granularity G are pairwise disjoint.
  Examples: Cities: $\mathit{Concepción} \wedge \mathit{Santiago} = \perp$
  Regions: $\mathit{Reg\_VIII} \wedge \mathit{Reg\_IX} = \perp$
- Granularity order is induced by granule order.
  - $G_1 \sqsubseteq G_2 \Leftrightarrow$
    $(\forall g_1 \in \mathrm{GrtoDom}_A(G_1))(\exists g_2 \in \mathrm{GrtoDom}_A(G_2))(g_1 \sqsubseteq g_2)$.

  Example: City $\sqsubseteq$ Region since every city is contained in some region.

The Granular Structure of Thematic Attributes

- Classification of multigranular attributes:

Place	Time	Births
Reg_I	*Q1Y2014*	n_1
Reg_II	*Q1Y2014*	n_2
...	...	...
Reg_XV	*Q1Y2014*	n_{15}

Thematic attributes: Usually numerical; typically on RHS of dependency.

Dimension attributes: Usually spatial or temporal; typically on LHS of dependency.

- Both thematic and dimension attributes have granular structure, although it arises and is used in different ways.
- The values of thematic attributes often involve imprecision.

General model: For each granularity, the numbers are partitioned into disjoint intervals.

Simple example: The intervals are defined by rounding.

  - One granularity for each i, $0 \leq i \leq r_{\text{max}}$. with granularity G_{round_i} corresponding to rounding to the nearest 10^i.

Granules: $g_1 \sqsubseteq g_2$ iff g_1 (as an interval) is contained in interval g_2.

Granularities: $G_1 \sqsubseteq G_2$ iff every interval (granule) associated with G_1 is contained in an interval (granule) associated with G_2.

Source database 1		
Place	Time	Births
Reg_I	*Q1Y2014*	n_1
Reg_II	*Q1Y2014*	n_2
...	...	...
Reg_XV	*Q1Y2014*	n_{15}

$$\sum_{i=1}^{15} n_i = b_1$$

Source database 2		
Place	Time	Births
Chile	*Q1Y2014*	b_1
Chile	*Q2Y2014*	b_2
Chile	*Q3Y2014*	b_3
Chile	*Q4Y2014*	null

- A constraint may involve a sum from on source equalling a value from a second.
- To formalize this, *aggregation operators* are defined on thematic attributes.
- These operators must be monotonic with respect to the thematic order.

Examples: summation, maximum

~~average~~ is not a valid aggregation operator because averaging is not monotonic in the required sense.

- Additional nonnegative numbers cannot decrease the sum but they can decrease the average.

Coarsening for Thematic Attributes

Coarsening maps a granule to the containing one of a coarser granularity.

- In this work, the use of coarsening is limited to thematic attributes.

Example:

$G_{I_{100}} =$ Intervals of the form $[n, n+99]$ with $n \geq 0$ divisible by 100.
$G_{I_{1000}} =$ Intervals of the form $[n, n+999]$ with $n \geq 0$ divisible by 1000.

  - Coarsen $\langle G_{I_{1000}}, [3100, 3199] \rangle = [3000, 3999]$.

- Coarsen $\langle G, g \rangle$ need not exist, but when it does, it is unique.

Principle: In general, for an aggregation operation to make sense, all operands must be of the same granularity.

Consequence: Coarsening must be applied to reduce operands to a common granularity.

Coarsening Tolerance for Thematic Attributes

Source database 1		
Place	Time	Births
Reg_I	*Q1Y2014*	n_1
Reg_II	*Q1Y2014*	n_2
...	...	...
Reg_XV	*Q1Y2014*	n_{15}

$$\sum_{i=1}^{15} n_i \leq b_1$$

Source database 2		
Place	Time	Births
Chile	*Q1Y2014*	b_1
Chile	*Q2Y2014*	b_2
Chile	*Q3Y2014*	b_3
Chile	*Q4Y2014*	null

- With data gathered from different sources, at different levels of aggregation, equality cannot be expected in general.
- The solution is to employ a *tolerance relation*.
- The values only need agree within a certain tolerance.
- The level of disagreement may depend upon the granularity of the thematic data.
- It may also depend upon the number of items in the aggregation.
- These ideas apply to inequality as well.

An Annotated Example TMCD

Source database 1		
Place	Time	Births
Reg_I	*Q1Y2014*	n_1
Reg_II	*Q1Y2014*	n_2
...	...	...
Reg_XV	*Q1Y2014*	n_{15}

$$\text{Place}\ \underline{\text{Time}} \xrightarrow[(\bot,1)]{=} \langle \text{Births}\colon \langle \theta, \textstyle\sum, \tau \rangle\rangle$$

$$\bigsqcup_{\text{Place}} \{Reg\_R \mid I \leq R \leq XV\} = Chile$$

Source database 2		
Place	Time	Births
Chile	*Q1Y2014*	b_1
Chile	*Q2Y2014*	b_2
Chile	*Q3Y2014*	b_3
Chile	*Q4Y2014*	null

$$(\forall T_1 \subseteq_f \text{Tuples}\langle\alpha\rangle)(\forall t_2 \in \text{Tuples}\langle\alpha\rangle)$$

Tuples of correct type

α = common relation type

$$(\forall G_1 \in \text{CoarsenSetMUB}_{\text{Births}}\langle\{t.\text{Births} \mid t \in T_1\}\rangle)$$
$$(\forall G_2 \in \text{GranSetOf}_{\text{Births}}\langle t_2.\text{Births}\rangle)$$
$$(\forall G \in \text{MUB}\langle\{G_1, G_2\}\rangle)$$

Find common granularity for birth values

$$((\bigwedge_{t_1 \in T_1} R\langle t_1\rangle) \wedge R\langle t_2\rangle$$

Tuples in relations

$$\wedge\ (\bigwedge_{t_1 \in T_1} (t_1.\text{Time} = t_2.\text{Time}))$$

Time value is the same in all tuples

T_1 = *Reg_i* tuples

t_2 = *Chile* tuple

$$\wedge\ ((\bigsqcup_{t_1 \in T_1}{}^{\bot}_{\text{Place}}\ t_1.\text{Place}) = t_2.\text{Place})$$

Place values match the governing rule

$$\Rightarrow \tau^{\langle G, \text{Card}(T_1)\rangle}_{\text{Births}} \langle \text{Coarsen}_{\text{Births}}\langle \sum^{G_1}_{t_1 \in T_1\ \text{Births}} \text{Coarsen}_{\text{Births}}\langle t_1.\text{Births}, G_1\rangle, G\rangle, \text{Coarsen}_{\text{Births}}\langle t_2.\text{Births}, G\rangle\rangle)$$

Aggregation tolerance

Coarsen sum to G

Aggregation at G_1

Coarsen *Reg_R* tuples to G_1

Coarsen *Chile* tuples to G

Conclusions and Further Directions

Conclusions:

Model for multigranular data: Extending the earlier work of Rodríguez and Bravo, and others, an extensive and formal model of multigranular attributes and relations has been developed.

TMCDs: Within this multigranular framework, *thematic multigranular comparison dependencies*, which recapture constraints which arise when data of differing granularities are to be integrated, have been developed.

Further Directions:

Data structures and algorithms: Although some initial ideas have been developed, it remains to develop detailed models for the data structures and algorithms which would underlie an efficient implementation.

Implementation and performance studies: A priority is to build a prototype system to test the ideas.

Elaboration of TMCDs: While TMCDs recapture common types of integration constraints, they are not complete. Further investigations are needed to identify other important types of constraints.