Classification of the Inconsistencies in the Databases Integration

Štuller, Július
1998

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

CLASSIFICATION OF THE INCONSISTENCIES IN THE DATABASES INTEGRATION

Július ŠTULLER

Technical report No. 746

May 1998

Institute of Computer Science, Academy of Sciences of the Czech Republic
Pod vodárenskou věží 2, 182 07 Prague 8, Czech Republic
phone: (+42 2) 6605 3200 fax: (+42 2) 85 85 789
e-mail: stuller@uivt.cas.cz
http://www.uivt.cas.cz/~stuller

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

CLASSIFICATION OF THE INCONSISTENCIES IN THE DATABASES INTEGRATION

Július ŠTULLER [1]

Technical report No. 746
May 1998

Abstract

The technological progress in the areas of the hardware and the software, together with the general expansion of the computers to almost all human activities, make it easier to realize the integration of many already existing databases.
Unfortunately the process of the integration of (existing) databases can be accompanied by many various difficulties and problems. One of them is surely the possible occurrence of the inconsistencies appearing in this process of the integration.
In the report we propose a classification of certain of these inconsistencies.

Keywords
database systems, logic, incomplete information

[1]This work was supported by the Grant No. 201/97/1070 of the Grant Agency of the Czech Republic : *Inconsistency Resolution Methods in the Data/Knowledge Base Integration* .

Contents

1 Introduction

The technological progress in the areas of the hardware, specially in the field of the (secondary) memories where the ever increasing capacities are paradoxaly in the last several years available at ever decreasing prices and smaller physical sizes, and the software, continuously more and more user friendly, efficient and cheaper, together with the general expansion of the computers to almost all human activities, make it easier to realize the integration of many already existing databases.
Every database can be seen, at least from the point of view of the logic, as a conjunction of different facts (and depending on the representation of these as data, information or knowledge, we can obtain either a classical database system, either an information system or even a kind of fashioned knowledge-base system) which leads naturally to the idea of representing such a database as a (formal) logic theory.
The states of such a database and the operations over such a database obey usually certain rules (so called integrity constraints in the database approach) which can again be expressed in the corresponding logic (for instance in the form of special axioms).
Unfortunately the process of the integration of (existing) databases (as an example see for instance the series of psycho-medical studies [Štuller, 1995a] & [Štuller, 1997a]) can be accompanied by many various difficulties and problems. One of them is surely the possible occurrence of the inconsistencies (in sense of the classical logic — see for instance [Štuller, 1995b] - [Štuller, 1995f]) appearing in this process of the integration of databases.

The inconsistencies in the integration of databases can occur at various levels and they can be of different types. In this paper we propose a certain classification of these inconsistencies.

2 Formulation of the Problem

We will study the problem of the inconsistencies in the integration of several databases under the following natural, at least from the point of view of the logic, assumption :

A1 : *Each of the considered databases to be integrated has no inconsistencies (when taken alone) .*

Furthermore, for reasons of the simplification, and having in mind the current situation in the area of the database technologies where the **Codd** *relational data model* prevail, we will suppose that :

A2 : *All the databases to be integrated are relational ones.*

3 The simplest case

Let $\mathcal{B}_i$, $i \in \{1,2\}$, be two databases, each consisting of one relation, say $R_i = \langle A_i , D_i , T_i \rangle$, $i \in \{1,2\}$, respectively.
(See Appendix for definitions and notations)

From all the usual relational operations (operators) the ones which can lead to any possible inconsistencies are :

- the *union* of the relations
- the *joins*
- and the corresponding *compositions*.

Let us start by the union (we will use its generalized form from [Štuller, 1997b]) :

Definition 1

Let $R_i = \langle A_i , D_i , T_i \rangle$ be m relations ($m \geq 2$) .
The $\boldsymbol{\pi}$ **- union** of relations R_i is the relation noted
$\bigcup_{\pi}{}_{i=1}^{m} \boldsymbol{R}_i = \langle \boldsymbol{A} , \boldsymbol{D} , \boldsymbol{T} \rangle$ such that :

1. $\boldsymbol{D}(\boldsymbol{A}) \cap (\bigcap_{i=1}^{m} \boldsymbol{D}_i (\boldsymbol{\pi}_i (\boldsymbol{A}_i))) \neq \emptyset$
2. $\boldsymbol{T} = \bigcup_{i=1}^{m} \boldsymbol{T}_i (\boldsymbol{\pi}_i (\boldsymbol{A}_i))$

Convention 1 In the case of permutations π_i being *identities* we will omit the prefix $\boldsymbol{\pi}$ **-** and speak shortly only about the **union** and note it $\bigcup_{i=1}^{m} \boldsymbol{R}_i$.

At first, again for the simplification, we will suppose relations $\boldsymbol{R}_i$ are defined over the same relational scheme $\boldsymbol{S} = \langle \boldsymbol{A} , \boldsymbol{D} \rangle$, that is : 1. $A_i = A$
2. $D_i = D$

Example 1

$\boldsymbol{R}_1$	
Name	**Position**
Peter	researcher

$\boldsymbol{R}_2$	
Name	**Position**
Peter	director

$\boldsymbol{R} = \boldsymbol{R}_1 \cup \boldsymbol{R}_2$	
Name	**Position**
Peter	researcher
Peter	director

Even in this very simple example without any further supplementary information it is impossible to decide whether an inconsistency appeared in the process of the integration of databases. Such a supplementary information is in general expressed in so called *integrity constraint(s).*

We will suppose that we have such an integrity constraint. Let it be the following:

Every value of the attribute **Name** *is associated with no more than one value of the attribute* **Position**. (A particular case of a so called functional dependency)

More formally :

$(\forall u, v \in T)((t(Name) = u(Name)) \Rightarrow (t(Position) = u(Position)))$

Let us denote by $\Im$ the set of all the possible integrity constraints over given universe of discourse $U = D(A)$ and by $\boldsymbol{I}$ a subset of the set $\Im$.

Let us further denote by $\boldsymbol{\mathcal{R}}(\boldsymbol{I})$ the set of all the relations over given universe of discourse satisfying $\boldsymbol{I}$. It is obvious that the following holds in general :

Lemma 1

$$((\boldsymbol{R}_1 \in \boldsymbol{\mathcal{R}}(\boldsymbol{I})) \wedge (\boldsymbol{R}_2 \in \boldsymbol{\mathcal{R}}(\boldsymbol{I}))) \not\Rightarrow ((\boldsymbol{R}_1 \cup \boldsymbol{R}_2) \in \boldsymbol{\mathcal{R}}(\boldsymbol{I}))$$

Returning to our example they are two possibilities :

- They are some erroneous data in at least one of the relations R_i :

  $(\exists\, i \in \widehat{m})(\exists\, R_i = \langle A_i, D_i, T_i \rangle)(\exists\, t \in T_i)(t \text{ is "incorrect"})$

  (i.e. this t does not represent correctly a fact from the reality we are trying to capture in a database - relation R_i ; in our Example it could mean that either Peter is not a researcher or that he is not a director ...)

- All data are correct but at least one of the integrity constraints is wrong :

  $(\exists\, i \in I)(i \text{ is "incorrect"})$

  (i.e. this i does not correctly reflect the reality we are trying to model ; in our Example it would mean that there could be more than one Position associated with one Name ...)

In both cases the incorrect items must be removed. Let us denote by :

- $\hat{\boldsymbol{R}}$ the subrelation of the relation $\boldsymbol{R}$ containing all "incorrect" data :
  $\hat{\boldsymbol{R}} = \langle A, D, \{ t : A \rightarrow D(A) \mid t \text{ is "incorrect"} \} \rangle$

- $\check{\boldsymbol{R}}$ the subrelation of the relation $\boldsymbol{R}$ containing no "incorrect" data :
  $\check{\boldsymbol{R}} = \boldsymbol{R} - \hat{\boldsymbol{R}}$

- $\hat{\boldsymbol{I}}$ the subset of the set $\boldsymbol{I}$ containing all "incorrect" integrity constraints :
  $\hat{\boldsymbol{I}} = \{ i \in I : i \text{ is "incorrect"} \}$

- $\check{\boldsymbol{I}}$ the subset of the set $\boldsymbol{I}$ containing no "incorrect" integrity constraints :
  $\check{\boldsymbol{I}} = \boldsymbol{I} - \hat{\boldsymbol{I}}$

In the first case, as the result of the "correction of the data", we obtain new relations (without erroneous data) $\check{\boldsymbol{R}}_k$ and the new union $\bigcup_{k=1}^{m} \check{\boldsymbol{R}}_k$.

In the second case, as the result of the "correction of the integrity constraints", we obtain new set of the integrity constraints $\check{I}$ (without wrong constraints) .

4 Generalization

Next we will suppose the relations $\boldsymbol{R}_k$ are defined over such different relational schema $\boldsymbol{S}_k = \langle\, \boldsymbol{A}_k \,,\, \boldsymbol{D}_k \,\rangle$ that there exists an appropriate permutation π in $|\, \widehat{A_k} \,|$ that the following holds : $\boldsymbol{D}_1 (\boldsymbol{A}_1) \cap \boldsymbol{D}_2 (\boldsymbol{\pi} (\boldsymbol{A}_2)) \neq \boldsymbol{\emptyset}$

Example 2

$\boldsymbol{R}_1$	
Name	**Position**
Peter	researcher

$\boldsymbol{R}_2$	
Name	**Function**
Peter	director

$\boldsymbol{R} = \boldsymbol{R}_1 \cup_\pi \boldsymbol{R}_2$	
Name	**Post**
Peter	researcher
Peter	director

The necessary prerequisite is the existence of the "appropriate" permutation π in $|\, \widehat{A_k} \,|$ which must be semantically justifiable for the concrete databases – relations :
In our *Example 2* we presuppose that the (names of the) attributes **Position** and **Function** are synonyms (i.e. they are semantically equivalent) .

If this is the case the same reasoning we used to the union of relations apply also to the π - union of the relations and so we can summarize :

Definition 2

Let R_k be m relations we want to make an union over ($m \geq 2$) ,
I_k be m corresponding sets of the integrity constraints and
I_{m+1} be the set of the integrity constraints corresponding
to the π - union of the relations R_k such that

$$\boldsymbol{I} = \bigcup_{k=1}^{m+1} \boldsymbol{I}_k \text{ is (logically) consistent.}$$

We will call the inconsistencies in the π - union $\bigcup_{\pi}{}_{k=1}^{m} \boldsymbol{R}_k$:

data inconsistencies $\Leftrightarrow$ $(\exists\, \boldsymbol{k} \in \widehat{\boldsymbol{m}}) (\check{\boldsymbol{R}}_k \neq \boldsymbol{R}_k)$
integrity constraints inconsistencies $\Leftrightarrow$ $(\exists\, \boldsymbol{k} \in \widehat{\boldsymbol{m}+1}) (\check{\boldsymbol{I}}_k \neq \boldsymbol{I}_k)$
semantical inconsistencies $\Leftrightarrow$ $(\exists\, \boldsymbol{k} \in \widehat{\boldsymbol{m}}) (\boldsymbol{\pi}_k \neq \boldsymbol{Identity})$

5 Joins

In the following we will study the properties of the joins in the process of the integration of the relational databases .

We will not need to study the properties of the corresponding compositions as they are expressible in joins (and projections).

Let us start by giving the definition of the simplest join (so called the *equi-join*) from [Štuller, 1997b] :

Definition 3

Let $R_i = \langle A_i , D_i , T_i \rangle$ be m relations ($m \geq 2$) and B_i be m sets of attributes such that

$$((B_i \subset A_i) (\forall i \in \widehat{m})) \wedge (\bigcap_{i=1}^{m} D_i (\pi_i (B_i)) \neq \emptyset) .$$

The *join* of the relations R_i , *according to the attributes sets* B_i , *with respect to the equality* , is the relation noted $*_{\pi_1(B_1)=\pi_i(B_i)} \boldsymbol{R}_i = \langle \boldsymbol{A} , \boldsymbol{D} , \boldsymbol{T} \rangle$ where:

$$(\boldsymbol{A} = \bigcup_{i=1}^{m} \boldsymbol{A}_i) \wedge (\boldsymbol{D} = \bigcup_{i=1}^{m} \boldsymbol{D}_i) \wedge (\boldsymbol{T} = *_{\pi_1 (B_1) = \pi_i (B_i)} \boldsymbol{T}_i)$$

Convention 2

In case of π_i being the identities, the equality of B_i and such that they are *maximal* (in *set inclusion* sense) with such a property , we will omit the *index* $_{\pi_1(B_1)=\pi_i(B_i)}$ by the $*$ and call the join the *natural join* of $\boldsymbol{R}_i$.

Example 3

$\boldsymbol{R}_1$	
Husband	**Wife**
Joseph	Mary

$\boldsymbol{R}_2$	
Mother	**Child**
Mary	Jesus

$\boldsymbol{R} = \boldsymbol{R}_1 * \boldsymbol{R}_2$		
Father	**Mother**	**Child**
Joseph	Mary	Jesus

Again, as in the case of the union, even in this very simple example without any further supplementary information it is impossible to decide whether an inconsistency appeared in the process of the integration of databases.

And, analogically to the case of the π - union , we could obtain the same three classes of the inconsistencies : *semantical*, *integrity constraints* and *data* inconsistencies.

6 Conclusion

By analyzing some simple problems we have arrived at the sources of possible inconsistencies in the integration of databases and we have proposed a certain classification of these inconsistencies based on the their sources.

The process of the integration of databases has been studied from mid-eighties, with the emphasis on the schema integration (see e.i. [Batini et al., 1986] as one of the first papers and [Ramesh & Ram, 1997] - [Santucci, 1998] as ones of the last ones) and with less attention on the integration of data themselves (see for instance [Orlowska et al., 1997]) . To our knowledge none of them proposed any kind of classification of the inconsistencies in the process of databases integration.

In the future we would like to further develop our classification of the inconsistencies in databases integration with the focus to the possibilities of designing some kind of support for the resolution of these inconsistencies.

7 Appendix

Definition

A **relation** in the RMD will be any triple $\langle\ A, D, T\ \rangle$ with

1. A being a finite set of **attribute names** .
2. D being a mapping which maps every attribute name $a \in A$ to a **domain** , noted $D(a)$.
   Let us : *denote* by $D(A)$ the *union* of all $D(a)$ and
   call it the **universe of discourse** .
3. T being a finite set of **mappings** t from A to the universe of discourse $D(A)$ such that $t(a) \in D(a)$ for all $a \in A$.

Notation 1

$$\widehat{\boldsymbol{m}} = \{1, 2, \cdots, m\} \qquad (\widehat{0} = \emptyset)$$

Notation 2

The *cardinality* of a set A will be denoted by $|\,\boldsymbol{A}\,|$.

Notation 3

Let $A_i = \{ a_{ij} \mid j \in \widehat{|A_i|} \}$, $i \in \{1, 2\}$.

$(\forall j \in \widehat{|A_i|})\ (D_1(a_{1j}) \cap D_2(a_{2\pi(j)}) \neq \emptyset\ )$

(π being an appropriate permutation in $\widehat{|A_i|}$)

$$\Updownarrow$$

$$\boldsymbol{D}_1(\boldsymbol{A}_1) \cap \boldsymbol{D}_2(\boldsymbol{\pi}(\boldsymbol{A}_2)) \neq \boldsymbol{\emptyset}$$

Lemma

$$((\boldsymbol{D}_1(\boldsymbol{A}_1) \cap \boldsymbol{D}_2(\boldsymbol{\pi}(\boldsymbol{A}_2))) \neq \boldsymbol{\emptyset}) \Rightarrow (|\,\boldsymbol{A}_1\,| = |\,\boldsymbol{A}_2\,|)$$

Notation 4

$$T = \{ t : A \to D(A) \mid (\exists i \in \widehat{m})(\exists u_i \in T_i)\ (u_i(\pi_i(A_i)) = t(A)) \}$$

$$\Updownarrow$$

$$\boldsymbol{T} = \bigcup_{i=1}^{m} \boldsymbol{T}_i(\boldsymbol{\pi}_i(\boldsymbol{A}_i))$$

Definition

Let $R = \langle\, A\,,\, D\,,\, T\,\rangle$ be a relation and $A_1 \subset A$.

The **projection** of the relation R over A_1 is the relation noted $\boldsymbol{R}\,[\,\boldsymbol{A}_1\,] = \langle\, A_1\,,\, D_1\,,\, T_1\,\rangle$ such that :

1. $\boldsymbol{D}_1 = \boldsymbol{D}/\boldsymbol{A}_1$
   (the *restriction* of the mapping D on the subset A_1 of A)
2. $\boldsymbol{T}_1 = \boldsymbol{T}\,[\,\boldsymbol{A}_1\,]$

Notation 5

$$T_1 = \{\, t : A_1 \rightarrow D_1(\,A_1\,) \mid (\,\exists\, u \in T\,)\,(\,t\,(\,A_1\,) = u\,(\,A_1\,)\,)\}$$
$$\Updownarrow$$
$$\boldsymbol{T}_1 = \boldsymbol{T}\,[\,\boldsymbol{A}_1\,]$$

Notation 6

$$D\,(\,a_j\,) = \bigcup_{i=1}^{m} D_i\,(\,a_j\,)\;,\quad \forall j \in |\,\widehat{A}\,| \qquad\qquad [\,\boldsymbol{A} = \bigcup_{i=1}^{m} \boldsymbol{A}_i\,]$$
$$\Updownarrow$$
$$\boldsymbol{D} = \bigcup_{i=1}^{m} \boldsymbol{D}_i$$

Definition

Let $R_i = \langle\, A_i\,,\, D_i\,,\, T_i\,\rangle,\ i \in \{\,1\,,\,2\,\}$, be two relations such that :

1. $(\,\exists\, \boldsymbol{A}_{21} \subset \boldsymbol{A}_2\,)\,(\,|\,\boldsymbol{A}_1\,| = |\,\boldsymbol{A}_{21}\,|\,)$
2. $\boldsymbol{D}_1\,(\,\boldsymbol{A}_1\,) \cap (\,\boldsymbol{D}_2\,/\boldsymbol{A}_{21})\,(\,\boldsymbol{\pi}\,(\,\boldsymbol{A}_{21}\,)\,) \neq \boldsymbol{\emptyset}\,)$
   (π being an appropriate permutation)
3. $\boldsymbol{T}_1\,(\,\boldsymbol{A}_1\,) \subset \boldsymbol{T}_2\,[\,\boldsymbol{A}_{21}\,]\,(\,\boldsymbol{\pi}\,(\,\boldsymbol{A}_{21}\,)\,)$

Then we will say that the relation R_1 is a **subrelation** of the relation R_2 — what we will note : $\boldsymbol{R}_1 \subset \boldsymbol{R}_2$.

Bibliography

[Batini et al., 1986] Batini C., Lenzerini M., Navathe S. B.: *A Comparative Analysis of Methodologies for Database Schema Integration.* ACM Computing Surveys, 1986, 18, 4, 323-364

[Orlowska et al., 1997] Orlowska M. E., Li H., Liu Ch.: *On Integration of Relational and Object-Oriented Database Systems.* LNCS 1338 (Proceedings of SOFSEM'97), 1997, 294-312

[Ramesh & Ram, 1997] Ramesh V., Ram S.: *Integrity Constraint Integration in Heterogeneous Databases: An Enhanced Methodology for Schema Integration.* Information Systems, 1997, 22, 8, 423-446

[Santucci, 1998] Santucci G.: *Semantic schema refinements for multilevel schema integration.* Data & Knowledge Engineering 1998, 25, 301-326

[Štuller, 1995a] Provazníková H., Štuller J., Štullerová N.: *Health and Prevention Care Possibilities for First Year Students of the Charles University 3rd Medical Faculty at Prague.* (In Czech: Zdraví a možnosti preventivní péče o studenty prvního ročníku 3. lékařské fakuty UK Praha.) In: Journal of Czechoslovak Psychology (*Československá psychologie*), 1995, 34/2, 159-169, Academia, ISSN 0009-062X

[Štuller, 1997a] Provazníková H., Štuller J., Štullerová N., Berkovičová V.: *Life Style, Health and Achievements of Students of the Faculty of Medicine.* (In Czech: Životní styl, zdraví a prospěch studentů lékařské fakulty.) In: Journal of Czechoslovak Psychology (*Československá psychologie*), 1997, 41/3, 216-224, Academia, ISSN 0009-062X

[Štuller, 1995b] Štuller J. : *Inconsistency Resolution in the Databases Integration.* In : Proceedings of the International Scientific Seminar DATABASE SYSTEMS ("DATABÁZOVÉ SYSTÉMY"), 1st-2nd June 1995, Bratislava, Slovakia, Centre of Edition, House of Technology of the Union of Slovak Scientific and Technical Societies, Bratislava, 102-107, ISBN 80-233-0348-1

[Štuller, 1995c] Štuller J.: *Inconsistency Problems in the Information Systems Integration.* (Invited Lecture.) In: BIOMATH-95 : Proceedings of the International Symposium on Mathematical Modelling and Information Systems in the Biology, Ecology and Medicine BIOMATH-95, 23rd-27th August 1995, Sofia, Bulgaria, (Editors : Popova E. D., Markov S., Ullrich Ch.), DATECS Publishing Ltd. - Institute of Biophysics, BAN, Sofia, 1995, 77, ISBN 954-613-005-2

[Štuller, 1995d] Štuller J.: *Inconsistency Conflict Resolution in the Integration of the Databases.* In: Proceedings of the International Conference "Computer Science", 5th-7th September 1995, Ostrava, Czech Republic, (Editors: Vondrák I., Štefan J.), Ostrava - MARQ, Ostrava Repronis, 283-289, ISBN 80-901751-7-1

[Štuller, 1995e] Štuller J.: *Inconsistency Resolution in the Integration of Databases.* In: Proceedings of the 15th Database Conference DATASEM '95, 8th-10th October 1995, Brno, Czech Republic, M. Š. - PRINT, Letovice, 47-52, ISBN 80-900066-9-8

[Štuller, 1995f] Štuller J.: *Inconsistency Conflict Resolution.* In: Proceedings of the XXII-th Winter School SOFSEM '95 - Seminar on Current Trends in Theory and Practice of Informatics, 25th November - 2nd December 1995, Milovy, Czech Republic, (Editors: Bartošek M., Staudek J., Wiedermann J.), Lecture Notes in Computer Sciences 1012, Springer-Verlag, Berlin, 1995, 469-474, ISBN 3-540-60609-2

[Štuller, 1997b] Štuller J.: *Database Systems and Logic - I.* Technical Report No. 702, ICS AS CR, Prague, 1997, 35 pages