Existence Conditions for the Inconsistencies in the Databases Integration

Štuller, Július
1998

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

EXISTENCE CONDITIONS FOR THE INCONSISTENCIES IN THE DATABASES INTEGRATION

Július ŠTULLER

Technical report No. 760

June 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

EXISTENCE CONDITIONS FOR THE INCONSISTENCIES IN THE DATABASES INTEGRATION

Július ŠTULLER [1]

Technical report No. 760
June 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, would it be in order to build up the very fashioned data warehousing, an enterprise wide data store, or simply one of many types of distributed data base systems providing different kind of the interoperability in the form of various heterogeneous, federated or multi-database systems.
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 study the existence conditions for 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 Formulation of the Problem

We will study the conditions for the existence of the inconsistencies in the integration of several databases under the following natural logical 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 :*

Let $\boldsymbol{\mathcal{B}}_i$, $i \in \widehat{m}$, be m *relational databases* to be integrated ($m \geq 2$) , each consisting of k_i *relations* $R^i{}_j = \langle A^i{}_j , D^i{}_j , T^i{}_j \rangle$.
(See Appendix for definitions and notations)

Let us denote by $\mathfrak{S}$ the set of all the possible *integrity constraints* over the given *universe of discourse* $\boldsymbol{\mathcal{U}} = \boldsymbol{\mathcal{D}} (\boldsymbol{A})$ where :

$$\boldsymbol{A} = \bigcup_{i=1}^{m} \bigcup_{j=1}^{k_i} A^i{}_j$$

and

$$\boldsymbol{\mathcal{U}} = \bigcup_{i=1}^{m} \bigcup_{j=1}^{k_i} D^i{}_j (A^i{}_j)$$

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 I .

We want to find the conditions which can lead to inconsistencies when trying to integrate some of the databases $\mathcal{B}_i$.

2 Integration by Unions of the Relations

We will suppose there exist (at least two) databases $\mathcal{B}_{i_1}$ and $\mathcal{B}_{i_2}$ each having (at least) one relation $R^{i_j}{}_{q_j}$ of the same cardinality , say c (greater or equal to two; if it was equal to one, the corresponding relations would be the lists, maybe ordered, which cannot lead to inconsistencies if the original relations had no inconsistencies ...) :

A3 : $(\exists\, c \geq 2)\,(\exists\, s \geq 2)\,(\forall\, j \in \hat{s})\,(\exists\, \mathcal{B}_{i_j})\,(\exists\, R^{i_j}{}_{q_j} \in \mathcal{B}_{i_j})\,(\mid A^{i_j}{}_{q_j} \mid = c)$

Remark : If all the databases $\mathcal{B}_i$ do not consist of (simple) lists, we can always find, by successive projections, corresponding (at least) couples of (sub)relations $R^{i_j}{}_{q_j}$ with the required properties.

First, for the simplification, we will suppose relations $R^{i_j}{}_{q_j}$ are defined over the same relational scheme $\boldsymbol{S} = \langle\, \boldsymbol{A}\,,\, \boldsymbol{D}\,\rangle$, that is : 1. $A^{i_j}{}_{q_j} = A$
2. $D^{i_j}{}_{q_j} = D$

A4 : $(\forall\, j \in \hat{s})\,(R^{i_j}{}_{q_j} \in S = \langle\, A\,,\, D\,\rangle)$

We have shown in [Štuller, 1998] that the following hold :

Lemma 1 : $((\exists \boldsymbol{s} \geq 2)(\forall \boldsymbol{j} \in \hat{\boldsymbol{s}})(\exists \boldsymbol{\mathcal{B}}_{i_j})(\exists \boldsymbol{R}^{i_j}{}_{q_j} \in (\boldsymbol{\mathcal{B}}_{i_j} \cap \boldsymbol{\mathcal{R}}(\boldsymbol{I}))) \not\Rightarrow (\bigcup_{j=1}^{s} \boldsymbol{R}^{i_j}{}_{q_j} \in \boldsymbol{\mathcal{R}}(\boldsymbol{I}))$

which can lead to inconsistencies (of two different types : *data inconsistencies* and *integrity constraints inconsistencies* - see [Štuller, 1998] for details) in the union.

But the union, to be meaningful, should be done only after a thorough semantical justification and verification because syntactical equality of attributes and of the corresponding domains may be misleading, especially in the case of overloaded concepts like *name*, *number*, *year* etc.

2.1 π - unions

We can relax the condition that the relations we want to make an union over are defined over the same relational scheme by requiring the existence of the permutations $\pi^{i_j}_{q_j}$ such that there exists the π - union of the relations ${R^{i_j}}_{q_j}$:

A5 : $$\bigcap_{i=1}^{s} {D^{i_j}}_{q_j} \, ({\pi^{i_j}}_{q_j} \, ({A^{i_j}}_{q_j} \,)) \neq \emptyset$$

If it is so, we can obtain the corresponding lemma for the π - union :

Lemma 2 :

$$((\exists \boldsymbol{s} \geq 2)(\forall \boldsymbol{j} \in \hat{\boldsymbol{s}})(\exists \mathcal{B}_{i_j})(\exists {\boldsymbol{R}^{i_j}}_{q_j} \in (\mathcal{B}_{i_j} \cap \mathcal{R}(\boldsymbol{I}))) \not\Rightarrow (\bigcup_{\pi}{}_{j=1}^{s} {\boldsymbol{R}^{i_j}}_{q_j} \in \mathcal{R}(\boldsymbol{I}))$$

which can again lead to some inconsistencies (apart from data and integrity constraints inconsistencies, mentioned before, to *semantical inconsistencies* - for details see again [Štuller, 1998]) .

Here, to obtain meaningful results, we must be even more careful to semantically justify the meaning of performing the operation of the π - union .

2.2 Resume

We have seen that the integration of databases by unions may lead to (different types of the) inconsistencies .

In order to eliminate as much as possible the occurrences of these inconsistencies one should try to, especially in the case of the validity of the conditions **A3 &**

- **A4 :** *clear the databases* to be integrated from :
  - **incorrect data** which *can lead to data inconsistencies*
  - **incorrect integrity constraints** which *can lead to integrity constraints inconsistencies*
- **A5 :** *semantically* **deeply analyze** the *corresponding attributes* in the relations to be integrated by π - unions . Missing to do so can lead to the *semantical inconsistencies*.

3 Integration by the Joins

We will illustrate the situation with the integration by joining the relations in the following examples .

Example 1

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

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

$\boldsymbol{R} = \boldsymbol{R}_1 *_{Wife=Mother} \boldsymbol{R}_2$		
Husband	**Wife**	**Child**
Joseph	Mary	Jesus

The comparison of the join with the π - union of the (same) relations :

$\boldsymbol{R} = \boldsymbol{R}_1 \bigcup_{\pi} \boldsymbol{R}_2$	
Man	**Woman**
Jesus	Mary
Joseph	Mary

shows that the integration by joins against the integration by unions :

- allows **new** *relationships between objects* (entities or their attributes) which
- can be the sources of **new** *semantical inconsistencies*
  (having for arguments some of such new relationships)
  in addition to the inconsistencies known from the unions .

Nevertheless, depending on the every concrete situation one must choose the best appropriate operation to perform the integration of the databases.

Example 2

$\boldsymbol{R}_1$	
Mother	**Son**
Eve	John

$\boldsymbol{R}_2$	
Mother	**Daughter**
Eve	Anne

$\boldsymbol{R} = \boldsymbol{R}_1 * \boldsymbol{R}_2$		
Mother	**Son**	**Daughter**
Eve	John	Anne

$\boldsymbol{R} = \boldsymbol{R}_1 \bigcup_{\pi} \boldsymbol{R}_2$	
Mother	**Child**
Eve	John
Eve	Anne

But in general what was said about the importance of the semantical justification for the π - union hold even more for the joins as the only condition on p relations ${R^{i_k}}_{q_k}$ to be joinable is :

A6 : $$\bigcap_{i=1}^{p} {D^{i_k}}_{q_k} ({\pi^{i_k}}_{q_k} ({B^{i_k}}_{q_k})) \neq \emptyset \qquad ({B^{i_k}}_{q_k} \subset {A^{i_k}}_{q_k}) (\forall k \in \widehat{p})$$

which is equal to the condition **A5** with the unique difference that ${B^{i_k}}_{q_k} \subset {A^{i_k}}_{q_k}$ and so one can have in principle $\prod_{i=1}^{p} 2^{|{A^{i_k}}_{q_k}|}$ possibilities of performing the join of p relations.

3.1 Resume

The integration of databases by joins may lead to similar types of the inconsistencies as in the case of the integration by unions.
In order to eliminate as much as possible the occurrences of these inconsistencies one should try to, especially in the case of the validity of the condition **A6**

- *clear the databases* to be integrated from :
  - **incorrect data** which *can lead to data inconsistencies*
  - **incorrect integrity constraints** which *can lead to integrity constraints inconsistencies*
- *semantically* **deeply analyze** the *corresponding attributes* in the relations to be integrated by joins .
  Missing to do so can lead to the *semantical inconsistencies.*

4 Other Operations of the Relational Algebra

All the other usual operations, except the compositions, do not contribute to the process of the integrations of databases. As concerns the different compositions, they are always expressible by joins (and projections) and so need not to be threated again.

5 Conclusion

We have find *four* conditions **A3**, **A4**, **A5** and **A6** under which can occur different types of the inconsistencies in the process of the integration of databases.
The conditions **A3**, **A4**, **A5** apply to the integration by unions while the conditions **A6** applies to the integrations by joins.

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, 1997]). To our knowledge none of them studied the existence conditions of the inconsistencies in the process of databases integration.

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

6 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** (of the relation) .
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

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

Notation 2

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

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 3

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

Notation 4

Let $A_i \ = \ \{\,a_{ij}\ |\ j \in \widehat{|\ A_i\ |}\,\}$, $i \in \{\,1,2\,\}$.
$(\,\forall\, j \in \widehat{|\ A_i\ |}\,)\ \ (\,D_1\,(\,a_{1j}\,)\ \cap\ D_2\,(\,a_{2\pi(j)}\,) \ \neq \ \emptyset\ )$
(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

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

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 : $\boldsymbol{R}_1 \ \subset \ \boldsymbol{R}_2$

Definition

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

Notation 5

$$\bigcup_{i=1}^{m} \boldsymbol{T}_i (\boldsymbol{\pi}_i (\boldsymbol{A}_i)) = \{ t : A \rightarrow D (A) \mid (\exists i \in \widehat{m}) (\exists u_i \in T_i) \\ (u_i (\pi_i (A_i)) = t (A)) \}$$

Convention

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 :
$\bigcup_{i=1}^{m} \boldsymbol{R}_i$.

Definition

Let $R_i = \langle A_i , D_i , T_i \rangle$ be m relations $(m \geq 2)$ and
B_i m sets of attributes such that : $B_i \subset A_i \quad (\forall i \in \widehat{m})$
$\bigcap_{i=1}^{m} D_i (\pi_i (B_i)) \neq \emptyset$.

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 :

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

Notation 6

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

$$\Updownarrow$$

$$\boldsymbol{D} = \bigcup_{i=1}^{m} \boldsymbol{D}_i$$

Notation 7

$$*_{\pi_1 (B_1) = \pi_i (B_i)} \boldsymbol{T}_i = \{ t : A \rightarrow D (A) \mid ((\forall i \in \widehat{m}) (\exists u_i \in T_i)) \\ ((t (A_j) = u_j (A_j)) \wedge \\ (u_1 (\pi_1 (B_1)) = u_i (\pi_i (B_i))))) \}$$

Convention

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 shortly the **natural join** of $\boldsymbol{R}_i$.

Convention

We will call a set of attributes a **compound attribute** or even, shortly, only an **attribute** .
When such a set will have *exactly one* element, we will call it, whenever necessary, a **simple attribute** .

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, 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

[Štuller, 1998] Štuller J.: *Classification of the Inconsistencies in the Databases Integration.* Technical Report No. 746, ICS AS CR, Prague, 1998, 10 pages