# Null values revisited in prospect of data integration

Guy de Tré, Rita de Caluwe, and Henri Prade*

Computer Science Laboratory, Department of Telecommunications and Information Processing, Ghent University, Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium

* Institut de Recherche en Informatique de Toulouse (IRIT)–CNRS, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex, France

**Abstract.** A considerable part of the information available in the real world is inherently imperfect and/or incomplete. Traditionally, the most commonly adopted modelling approaches for dealing with imperfect and missing information are based on the use of default values and of null values. However, in order to deal with imperfect information in a more efficient way, a database system needs some more advanced modelling facilities that better reflect the semantics of the data. Such facilities become even more necessary if data stemming from different data sources must be integrated in a distributed, federated database system. In this paper, the concept of a 'null' value has been revisited. A semantically richer definition, based on possibility theory, is proposed together with a description of the accompanying many-valued logic. Additionally, the potentials of the new approach with respect to the integration of data in multi-database environments or grid database services are illustrated.
**keywords:** Null values, many-valued logic, data semantics, database modelling, data integration.

## 1 Introduction

The treatment of missing information in traditional database models has been widely addressed in research and continues to be explored. A survey that gives an overview of the field is presented in [13]. The most commonly adopted technique is to model missing data with a pseudo-description, called *null*, that denotes 'missing' [4, 2, 5]. Once 'null' values are admitted into a database, it is necessary to define the impact of transformations and modifications in the presence of 'nulls'. On the other hand, techniques for the handling of imperfect information have been studied by the 'fuzzy' database community [3, 17, 9, 16] and research on such techniques has been recognized as a challenge for the near future [15].

In his approach, Codd has extended the relational calculus based on an underlying three-valued logic [4, 5] in order to formalize the semantics of 'null' values in traditional databases. Mainly due to the fact that the law of excluded middle no longer holds, this extension has been subject to criticism [7]. As an alternative approach, Date proposes to omit null values and to force the database

administrator or some suitable authorized user to select a specific, so-called 'default' value from the domain of the field of the missing data to denote missing information in that field [7, 8].

In this paper, it is shown how possibility theory can help to overcome some of the problems encountered in a traditional 'null' value approach and moreover allows to deal efficiently with both imperfect and missing information. Furthermore, it is discussed how the presented approach supports the integration of (imperfect) data stemming from different data sources. In Section 2 a new approach for dealing with 'null' values in databases is presented. In order to define the full semantics of the new concepts a many-valued logic, based on so-called extended possibilistic truth values (EPTV's), has been used. An overview of some basic definitions and properties of EPTV's is given in Section 3. In Section 4 it is illustrated how the presented approach can contribute to semantically richer data integration techniques. Finally, some concluding remarks are given in Section 5.

## 2 Null values in databases

### 2.1 Traditional approaches

Missing information in databases could be indicated and handled by using a 'null' value, which can be seen as a special mark that denotes the fact that the actual database value is missing [2, 5]. In order to assign correct semantics to such a mark, it is important to distinguish between two main reasons for information being missing. As originally stated by Codd [5], information is either missing because:

- data is unknown to the users, but the data is applicable and can be entered whenever it happens to be forthcoming;
- data is missing because it pertains to a property that is inapplicable to the particular object represented by the database instance involved.

Accordingly, Codd introduces the idea of making an explicit distinction in the modelling of both cases by using two different kinds of null values, one meaning "value unknown" and the other "value not defined" [6]. Nevertheless, in many traditional approaches (and existing database systems) such an explicit distinction is not made and a single kind of 'null' values is used to handle both cases of missing data.

In formal definitions of database models null values are represented by some special symbol, e.g. by the bottom symbol '$\perp$' [1, 20]. In some formal approaches, null values are considered to be domain dependent [20]: the domain $dom_t$ of each data type $t$ supported by the database model contains a domain specific null value $\perp_t$, which implies that an explicit distinction is made between for example a missing integer value, a missing string value, etc.

In order to define the impact of transformations and modifications in the presence of null values, a many-valued logic [19] has been used. This logic is

three-valued if only a single kind of null values is used [2, 5] and four-valued if two distinct kinds of null values are considered [6]. The truth values of Codd's four-valued logic are resp. true ($T$), false ($F$), unknown ($\perp_U$) and inapplicable ($\perp_I$). In Codd's three-valued logic, the latter two values have been combined into one truth value $\perp_{U/I}$, which stands for 'either unknown or inapplicable'.

### 2.2 Problem description

A problem with many-valued logics is that the law of excluded middle and the law of non-contradiction do not hold. For example, in a three-valued Kleene logic the considered truth values are $T$, $F$ and $\perp$, conjunction is defined by $T \wedge T = T$, $T \wedge F = F \wedge T = F \wedge F = F \wedge \perp = \perp \wedge F = F$ and $T \wedge \perp = \perp \wedge T = \perp \wedge \perp = \perp$; disjunction by $T \vee F = F \vee T = T \vee T = T \vee \perp = \perp \vee T = T$, $F \vee F = F$ and $F \vee \perp = \perp \vee F = \perp \vee \perp = \perp$; and negation by $\neg T = F$, $\neg F = T$ and $\neg \perp = \perp$ [19]. With these definitions $\perp \wedge \neg(\perp) = \perp \neq F$ and $\perp \vee \neg(\perp) = \perp \neq T$.

Another important observation is that considering a truth value that is interpreted as 'unknown' (as e.g. $\perp_U$ or $\perp_{U/I}$), induces a problem of truth functionality because the 'degree of' truth of a proposition can no longer be calculated from the 'degrees of truth' of its constituents. 'Unknown' stands for the *uncertainty* of whether a proposition is true or false, which differs from the idea of 'many-valuedness' in a logical format where many-valued logics are intended for: degrees of uncertainty and degrees of truth are different concepts [12].

These observations explain some of the rationales behind Date's criticism on the use of null values [7]. In the approach, presented in the remainder of the paper, unknown information —more specifically uncertainty about the value of existing information— is no longer modelled using a special 'null' value, but using possibility theory, which is intended to model uncertainty [22, 11].

### 2.3 Revised approach

In the revised approach, only one kind of 'null' values is considered: 'null' values that represent 'inapplicability' and denote that a regular domain value is not applicable. In accordance with [20], these 'null' values are considered to be domain dependent. Therefore, the domain $dom_t$ of each data type $t$ supported by the database model is considered to contain such a null value $\perp_t$.

An unknown value for an attribute $A$ with associated type $t$ is modelled by a possibility distribution $\pi_A$ that is defined over the domain $dom_t$ of $t$ [22, 11]. A possibility distribution can be derived from a fuzzy set, which on its turn is a generalization of a set. Each fuzzy set $\tilde{V}$ is characterized by a membership function

$$\mu_{\tilde{V}} : dom_t \rightarrow [0, 1] : x \mapsto \mu_{\tilde{V}}(x)$$

which associates a membership grade $\mu_{\tilde{V}}(x) \in [0, 1]$ with each element of the universe over which the fuzzy set is defined. A membership grade 0 denotes that $x$ does not belong to the fuzzy set, a membership grade 1 denotes that $x$ completely belongs to the fuzzy set, whereas a membership grade $\mu_{\tilde{V}}(x) \in ]0, 1[$

denotes that $x$ only belongs to the set to a given extent. When representing a possibility distribution, a fuzzy set is interpreted as being disjunctive where each of its elements is a possible value and its associated membership grade denotes its degree of possibility.

For example, considering an attribute salary of an employee record, the values 'moderate' and 'unknown' can be defined by possibility distributions as given in Figure 1.

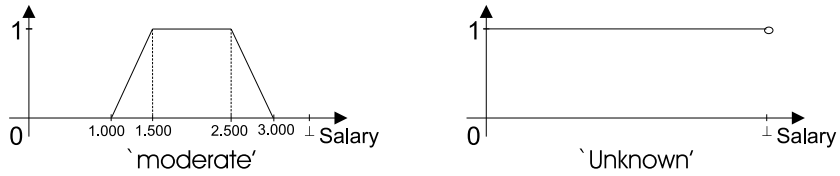


**Fig. 1.** Possibility distributions representing a moderate salary and an unknown salary.

This approach for modelling unknown information, is in fact the approach taken in so-called possibilistic databases as originally presented in [18]. When applied in a traditional database model, one can use the three label values 'UNK', 'N/A' and 'UNA' to represent 'unknown', 'not applicable' and 'unavailable' data. Thus, 'UNA' stands for the case where it is not even known if a considered property applies or not. The semantics of these labels is defined in terms of a possibility distribution:

- 'UNK' corresponds to the possibility distribution with membership function

$$\begin{aligned}\mu_{UNK}(x) &= 1, \text{ if } x \in dom_t \setminus \{\perp_t\} \\ &= 0, \text{ if } x = \perp_t,\end{aligned}$$

- 'N/A' corresponds to the possibility distribution with membership function

$$\begin{aligned}\mu_{N/A}(x) &= 0, \text{ if } x \in dom_t \setminus \{\perp_t\} \\ &= 1, \text{ if } x = \perp_t \text{ and}\end{aligned}$$

- 'UNA' and corresponds to the possibility distribution with membership function

$$\mu_{UNA}(x) = 1, \forall\, x \in dom_t.$$

Thus, beside $\perp_t$ no additional domain values are required and 'unknown' and 'unavailable' stand for uncertainty, which is conform with [12].

## 3 Extended possibilistic truth values

An adequate logic to support the revised null value concept has been obtained by imposing the same possibilistic uncertainty to a three-valued logic with truth

values true ($T$), false ($F$) and inapplicable ($\perp$). Such a logic, based on a three-valued Kleene logic, has been developed in [10]. The resulting truth values have been called 'extended possibilistic truth values' (EPTV's). EPTV's provide an epistemological representation of the truth of a proposition, which allows to reflect knowledge about the actual truth. Their semantics is defined in terms of a possibility distribution.

**Definition 1 (EPTV)** *With the understanding that $P$ represents the universe of all propositions and $\tilde{\wp}(I^*)$ denotes the set of all ordinary fuzzy sets in the universe $I^* = \{T, F, \perp\}$, the so-called extended possibilistic truth value (EPTV) $\tilde{t}^*(p)$ of a proposition $p \in P$ is formally defined by the mapping*

$$\tilde{t}^* : P \rightarrow \tilde{\wp}(I^*)$$

*which associates a fuzzy set $\tilde{t}^*(p)$ with each $p \in P$. The semantics of this associated fuzzy set are defined in terms of a possibility distribution.*
□

The EPTV $\tilde{t}^*(p)$ of a proposition $p$ can thus be seen as a possibility distribution

$$\tilde{t}^*(p) = \{(T, \mu_T), (F, \mu_F), (\perp, \mu_\perp)\}, \text{ where } \mu_T, \mu_F, \mu_\perp \in [0, 1]$$

Hereby, the membership grade $\mu_T$ denotes the possibility that $p$ is true, $\mu_F$ denotes the possibility that $p$ is false and $\mu_\perp$ denotes the possibility that some of) the elements of $p$ are not applicable, undefined or not supplied.

Special cases of EPTV's are:

| $\tilde{t}^*(p)$ | Interpretation |
|---|---|
| $\{(T, 1)\}$ | $p$ is true |
| $\{(F, 1)\}$ | $p$ is false |
| $\{(T, 1), (F, 1)\}$ | $p$ is unknown |
| $\{(\perp, 1)\}$ | $p$ is inapplicable |
| $\{(T, 1), (F, 1), (\perp, 1)\}$ | information about $p$ is unavailable |

These cases are verified as follows:

- If it is completely possible that the proposition is true and no other truth values are possible, then the proposition is considered to be true.
- If it is completely possible that the proposition is false and no other truth values are possible, then the proposition is considered to be false.
- If it is completely possible that the proposition is true, it is completely possible that the proposition is false and it is not possible that the proposition is inapplicable, then the proposition is considered to be applicable, but unknown. This truth value will shortly be called 'unknown'.
- If it is completely possible that the proposition is inapplicable and no other truth values are possible, then the proposition is considered to be inapplicable.

– If all truth values are completely possible, then this means that no information about the truth of the proposition is available. The proposition might be inapplicable, but might also be true, false or unknown. This truth value will shortly be called 'unavailable'.

New propositions can be constructed from existing propositions, using logical operators. An unary operator '$\tilde{\neg}$' is provided for the negation ($NOT$) of a proposition and binary operators '$\tilde{\wedge}$', '$\tilde{\vee}$', '$\tilde{\Rightarrow}$' and '$\tilde{\Leftrightarrow}$' are respectively provided for the conjunction ($AND$), disjunction ($OR$), implication ($IF\ THEN$) and equivalence ($IFF$) of propositions. The arithmetic rules to calculate the EPTV of a composite proposition and the algebraic properties of EPTV's are presented in [10]. The rules for negation, conjunction and disjunction can be summarized as

– *Rule for negation:* $\forall\, p \in P : \tilde{t}^*(NOT\ p) = \tilde{\neg}(\tilde{t}^*(p))$ where

$$\tilde{\neg} : \tilde{\wp}(I^*) \to \tilde{\wp}(I^*) : \tilde{V} \mapsto \tilde{\neg}(\tilde{V})$$

is defined by
- $\mu_{\tilde{\neg}(\tilde{V})}(T) = \mu_{\tilde{V}}(F)$
- $\mu_{\tilde{\neg}(\tilde{V})}(F) = \mu_{\tilde{V}}(T)$
- $\mu_{\tilde{\neg}(\tilde{V})}(\bot) = \mu_{\tilde{V}}(\bot)$

– *Rule for conjunction:* $\forall\, p, q \in P : \tilde{t}^*(p\ AND\ \ q) = \tilde{t}^*(p) \tilde{\wedge} \tilde{t}^*(q)$ where

$$\tilde{\wedge} : \tilde{\wp}(I^*) \times \tilde{\wp}(I^*) \to \tilde{\wp}(I^*) : (\tilde{U}, \tilde{V}) \mapsto \tilde{U}\ \tilde{\wedge}\ \tilde{V}$$

is defined by
- $\mu_{\tilde{U} \tilde{\wedge} \tilde{V}}(T) = \min(\mu_{\tilde{U}}(T), \mu_{\tilde{V}}(T))$
- $\mu_{\tilde{U} \tilde{\wedge} \tilde{V}}(F) = \max \begin{pmatrix} \min(\mu_{\tilde{U}}(T), \mu_{\tilde{V}}(F)), \\ \min(\mu_{\tilde{U}}(F), \mu_{\tilde{V}}(T)), \\ \min(\mu_{\tilde{U}}(F), \mu_{\tilde{V}}(F)), \\ \min(\mu_{\tilde{U}}(F), \mu_{\tilde{V}}(\bot)), \\ \min(\mu_{\tilde{U}}(\bot), \mu_{\tilde{V}}(F)) \end{pmatrix}$
- $\mu_{\tilde{U} \tilde{\wedge} \tilde{V}}(\bot) = \max \begin{pmatrix} \min(\mu_{\tilde{U}}(T), \mu_{\tilde{V}}(\bot)), \\ \min(\mu_{\tilde{U}}(\bot), \mu_{\tilde{V}}(T)), \\ \min(\mu_{\tilde{U}}(\bot), \mu_{\tilde{V}}(\bot)) \end{pmatrix}$

– *Rule for disjunction:* $\forall\, p, q \in P : \tilde{t}^*(p\ OR\ \ q) = \tilde{t}^*(p) \tilde{\vee} \tilde{t}^*(q)$ where

$$\tilde{\vee} : \tilde{\wp}(I^*) \times \tilde{\wp}(I^*) \to \tilde{\wp}(I^*) : (\tilde{U}, \tilde{V}) \mapsto \tilde{U}\ \tilde{\vee}\ \tilde{V}$$

is defined by
- $\mu_{\tilde{U} \tilde{\vee} \tilde{V}}(T) = \max \begin{pmatrix} \min(\mu_{\tilde{U}}(T), \mu_{\tilde{V}}(T)), \\ \min(\mu_{\tilde{U}}(T), \mu_{\tilde{V}}(F)), \\ \min(\mu_{\tilde{U}}(T), \mu_{\tilde{V}}(\bot)), \\ \min(\mu_{\tilde{U}}(F), \mu_{\tilde{V}}(T)), \\ \min(\mu_{\tilde{U}}(\bot), \mu_{\tilde{V}}(T)) \end{pmatrix}$
- $\mu_{\tilde{U} \tilde{\vee} \tilde{V}}(F) = \min(\mu_{\tilde{U}}(F), \mu_{\tilde{V}}(F))$

- $\mu_{\tilde{U}\tilde{\vee}\tilde{V}}(\perp) = \max \begin{pmatrix} \min(\mu_{\tilde{U}}(F), \mu_{\tilde{V}}(\perp)), \\ \min(\mu_{\tilde{U}}(\perp), \mu_{\tilde{V}}(F)), \\ \min(\mu_{\tilde{U}}(\perp), \mu_{\tilde{V}}(\perp)) \end{pmatrix}$

These rules are obtained by applying Zadeh's extension principle [21] to the operators of the three-valued Kleene logic [19]. Kleene logics are truth-functional, which means that according to these systems, the behavior of a logical operator is mirrored in a logical function combining Kleene truth values. Therefore, the extended truth value of every composed proposition can be calculated as a function of the extended truth values of its original propositions.

In the framework EPTV's the concept 'unknown', i.e. uncertainty about a truth value is not modelled by means of an extra value, but by means of possibility theory. An extra 'null' value is no longer needed for the modelling of incompleteness due to unavailability. However, an extra 'null' value remains necessary for the modelling of incompleteness due to inapplicability.

Using a special 'null' value to handle inapplicable information brings along with it the same problem as incomplete truth-functionality [7]. The point about this problem is that a database "is" not the real world, but instead contains only (partial) knowledge about the real world. By using extended possibilistic truth values this is reflected by the inability to cope adequately with the special cases $\tilde{t}^*(p\ AND\ NOT\ p)$ and $\tilde{t}^*(p\ OR\ NOT\ p)$. In the presented approach, this inability has been accepted as a tradeoff for being able to cope explicitly with the inapplicability of information.

However, this could be solved by using a constrained version of the extension principle, where one enforces constraints expressing

i. that the truth values of $p$ and $NOT\ p$ are a function of each other, and
ii. the piece of information, when available, that the truth value of $p$ is different from (or is equal to) $\perp$ (since we should have $\tilde{t}^*(p\ AND\ NOT\ p) = F$, $\tilde{t}^*(p\ OR\ NOT\ p) = T$, in case of applicability of the property, and $\tilde{t}^*(p\ AND\ NOT\ p) = \perp = \tilde{t}^*(p\ OR\ NOT\ p)$ otherwise).

## 4 Data integration in a multi-database environment

The approach presented in the previous sections is able to support a semantically richer handling of missing information in databases and is moreover consistent with the data modelling techniques for imperfect information used in possibilistic databases. Additionally, the approach allows for semantically richer techniques for combining and integrating data that stem from different (heterogeneous) data sources as for example is the case in data warehouses and grid databases.

In order to illustrate the advantages and potentials of the revised 'null' value approach for data integration, two different data sources $A$ and $B$ are considered. Furthermore, for the sake of the example it is assumed that each source contains, among other things, data about employee salaries. The salary attributes in both sources are respectively denoted by $salary_A$ and $salary_B$, whereas a value of

$salary_A$ is denoted by $v_A$ and a value of $salary_B$ is denoted by $v_B$. Further assume that it is necessary to integrate the data from both sources and that hereby a record value $v_A$ must be combined with a record value $v_B$ (e.g. because both records contain information about the same entity). Ten different cases could occur, all of them are described below. For each case, a comparison has been made between a *traditional approach* —where missing information is handled by one single null value, denoted by $NULL$—, the *revised null value approach* applied on a traditional data model as described in this paper and the revised null value approach applied on a *possibilistic data model*, which allows to model imprecise values and thus provides more facilities for the modelling of imperfect information.

*Case 1. Both $v_A$ and $v_B$ are regular values.*

- *Traditional approach* and *revised null value approach.*
  - If $v_A = v_B$ then both values can be combined with resulting value $v = v_A = v_B$.
  - If $v_A \neq v_B$ then the data of both sources are conflicting and the necessary precautions must be taken to prevent an inconsistent database.
- *Possibilistic approach.*
  Hereby, $v_A$ and $v_B$ are possibility distributions with membership functions $\mu_{v_A}$ and $\mu_{v_B}$ as described in Section 2. In order to compare two possibility distributions in prospect of integration the support function *supp* can be used. This function is defined by $supp(v) = \{x | x \in U \wedge \mu_v(x) > 0\}$ —where $U$ is the universe over which the possibility distribution $v$ is defined— and results in the set of all values of $U$ that are considered to be possible within $\tilde{v}$.
  - If $supp(v_A) \cap supp(v_B) \neq \emptyset$, then $v_A$ and $v_B$ can be combined by applying a conjunction operator (t-norm operator [14]) for fuzzy sets. A commonly used t-norm operator is the min-operator, with which the membership function $\mu_v$ of the resulting possibility distribution $v$ becomes $\mu_v = \min(\mu_{v_A}, \mu_{v_B})$.
  - If $supp(v_A) \cap supp(v_B) = \emptyset$ then the data of both sources are conflicting and again the necessary precautions must be taken to prevent inconsistency.

*Case 2. One of the values, e.g. $v_A$, is a regular value and the other one is 'unknown'.*

- *Traditional approach.*
  In this case $v_B = NULL$. Since $NULL$ can stand for both 'unknown' and 'inapplicable' this case can not be handled correctly without additional information.
- *Revised null value approach.*
  In this case $v_B = UNK$, as described in Section 2. The resulting value $v$ equals $v_A$, which is verified by the fact that $\min(\mu_{v_A}, \mu_{UNK}) = \mu_{v_A}$, with $\mu_{v_A}$ being the membership function of the possibility distribution that corresponds to $v_A$ and is defined by $\mu_{v_A}(v_A) = 1$ and $\mu_{v_A}(v) = 0$, if $v \neq v_A$.

– *Possibilistic approach.*
Hereby, both $v_A$ and $v_B = UNK$ are possibility distributions. This case can be identically handled as the possibilistic approach where both $v_A$ and $v_B$ are regular values. Due to the definition of 'UNK', $supp(v_A) \cap supp(UNK) \neq \emptyset$ and the resulting possibility distribution $v$ will equal $\mu_v = \min(\mu_{v_A}, \mu_{UNK}) = \mu_{v_A}$.

*Case 3. One of the values, e.g. $v_A$, is a regular value and the other one is 'inapplicable'.*

– *Traditional approach.*
In this case $v_B = NULL$. Since $NULL$ can stand for both 'unknown' and 'inapplicable' this case can not be handled correctly without additional information.
– *Revised null value approach.*
In this case $v_B = N/A$. The data of both sources are conflicting.
– *Possibilistic approach.*
Both $v_A$ and $v_B = N/A$ are possibility distributions. However, $supp(v_A) \cap supp(v_B) = \emptyset$, which denotes that the data of both sources are conflicting.

*Case 4. One of the values is a regular value, e.g. $v_A$, and the other one is 'unavailable'.*

– *Traditional approach.*
Again $v_B = NULL$, which can stand for both 'unknown' and 'inapplicable'. More information is necessary for a correct combination of the data.
– *Revised null value approach.*
In this case $v_B = UNA$. The resulting value $v$ equals $v_A$, which is verified by the fact that $\min(\mu_{v_A}, \mu_{UNA}) = \mu_{v_A}$, with $\mu_{v_A}$ being the membership function of the possibility distribution that corresponds to $v_A$ and is defined by $\mu_{v_A}(v_A) = 1$ and $\mu_{v_A}(v) = 0$, if $v \neq v_A$.
– *Possibilistic approach.*
Hereby, both $v_A$ and $v_B = UNA$ are possibility distributions and can be handled as in Case 1. With the definition of 'UNA', $supp(v_A) \cap supp(UNA) \neq \emptyset$ so that the resulting possibility distribution $v$ equals $\mu_v = \min(\mu_{v_A}, \mu_{UNA}) = \mu_{v_A}$.

*Case 5. Both $v_A$ and $v_B$ are 'unknown'.*

– *Traditional approach.*
In this case $v_A = v_B = NULL$. 'NULL' can stand as well for 'unknown' information, as for 'inapplicability'. Additional information that clarifies the correct interpretations of the 'NULLs', is required for a correct handling.
– *Revised null value approach* and *possibilistic approach.*
In both approaches $v_A = v_B = UNK$. The resulting value is $v = v_A = v_B = UNK$, which is verified by $\min(\mu_{UNK}, \mu_{UNK}) = \mu_{UNK}$.

*Case 6. One of the values, e.g. $v_A$, is 'unknown' and the other one is 'inapplicable'.*

- *Traditional approach.*
  Again $v_A = v_B = NULL$. Additional information that clarifies the correct interpretations of the 'NULLs', is required for a correct handling.
- *Revised null value approach* and *possibilistic approach.*
  In both approaches $v_A = UNK$ and $v_B = N/A$. The data of both sources are conflicting since $supp(UNK) \cap supp(N/A) = \emptyset$ and the necessary precautions must be taken.

*Case 7. One of the values, e.g. $v_A$, is 'unknown' and the other one is 'unavailable'.*

- *Traditional approach.*
  $v_A = v_B = NULL$. Additional information is required for a correct handling.
- *Revised null value approach* and *possibilistic approach.*
  In both approaches $v_A = UNK$ and $v_B = UNA$. With the definitions of 'UNK' and 'UNA', $supp(UNK) \cap supp(UNA) \neq \emptyset$ and the resulting value $v = UNK$ ($\mu_v = \min(\mu_{UNK}, \mu_{UNA}) = \mu_{UNK}$).

*Case 8. Both $v_A$ and $v_B$ are 'inapplicable'.*

- *Traditional approach.*
  $v_A = v_B = NULL$. Additional information is required for a correct handling.
- *Revised null value approach* and *possibilistic approach.*
  In both approaches $v_A = v_B = N/A$. The resulting value is $v = v_A = v_B = N/A$, which is verified by $\min(\mu_{N/A}, \mu_{N/A}) = \mu_{N/A}$.

*Case 9. One of the values, e.g. $v_A$, is 'inapplicable' and the other one is 'unavailable'.*

- *Traditional approach.*
  $v_A = v_B = NULL$. Additional information is required for a correct handling.
- *Revised null value approach* and *possibilistic approach.*
  In both approaches $v_A = N/A$ and $v_B = UNA$. The resulting value is $v = v_A = N/A$, which is verified by $\min(\mu_{N/A}, \mu_{UNA}) = \mu_{N/A}$.

*Case 10. Both $v_A$ and $v_B$ are 'unavailable'.*

- *Traditional approach.*
  $v_A = v_B = NULL$. This case is correctly handled since $NULL$ traditionally stands for $UNK$ of $N/A$
- *Revised null value approach* and *possibilistic approach.*
  In both approaches $v_A = v_B = UNA$. The resulting value is $v = v_A = v_B = UNA$, which is verified by $\min(\mu_{UNA}, \mu_{UNA}) = \mu_{UNA}$.

For the cases of conflict encountered in cases 1, 3, 6, in case imprecise information would be allowed in the database, one may store the disjunction of the conflicting pieces f information.

With the previous case study it is illustrated how the presented approach for the modelling of missing information can contribute to semantically richer data integration and combination techniques and therefore can also contribute to semantically richer grid database services. Compared with a traditional approach where one 'null' value is considered —as is the case in most database models—, the revised null value approach provides a semantically richer modelling technique. At the same time the revised null value approach is completely consistent with the possibilistic data modelling approach and can therefore be easily integrated in any possibilistic database model.

Compared with an approach where two distinct kinds of null values are used to make a distinction between "value unknown" and "value not defined" (as proposed in [6]), the revised approach presented in this paper has the advantage that the case "value unknown" no longer requires an extra truth value in the underlying logic.

## 5 Conclusion

In this paper the semantics of null values has been revised. A semantically richer approach, based on possibility theory and an accompanying logic on extended possibilistic truth values, has been proposed. Furthermore, the usefulness of the new approach has been illustrated in prospect of the need for semantically richer data integration and data combination techniques as might be required in heterogeneous multi-database environments and advanced grid database services.

By explicitly making a distinction between uncertainty and truth, whereby 'unknown' has been modelled as standing for the uncertainty of whether a proposition is true or false, an extra truth (and 'null') value for the representation of 'unknown' has been avoided. Consequently, the problems connected with the existence of such an extra truth value, can be solved.

However, in order to be able to cope with information that is missing due to inapplicability or non-existence of information an extra truth value $\perp$ has been introduced, as well as the incorporation of an extra domain specific element $\perp_t$ in the domains $dom_t$ of the data types $t$ supported by the database model. This extra truth value brings along with it the same problem as incomplete truth-functionality. This is reflected by the inability to cope adequately with the special cases $\tilde{t}^*(p\ AND\ NOT\ p)$ and $\tilde{t}^*(p\ OR\ NOT\ p)$. This inability has been accepted in the paper, as a tradeoff for being able to cope explicitly with the inapplicability of information.

## References

1. Abiteboul, S., Hull, R., Vianu, V.: Foundations of databases. Addison-Wesley Publishing Company, Reading, USA (1995).

2. Biskup, J.: A Formal Approach to Null Values in Database Relations. In: Advances in Data Base Theory. Gallaire H., Minker J., Nicolas J. (eds.), Plenum Press, New York, USA (1981) 299–341.
3. Bosc P., Kacprzyk J. (eds.): Fuzziness in Database Management Systems. Physica-Verlag, Heidelberg, Germany (1995).
4. Codd, E.F.: RM/T: Extending the Relational Model to capture more meaning. ACM Transactions on Database Systems **4** 4 (1979).
5. Codd, E.F.: Missing Information (Applicable and Inapplicable) in Relational Databases. ACM SIGMOD Record **15** 4 (1986) 53–78.
6. Codd, E.F.: More Commentary on Missing Information in Relational Databases (Applicable and Inapplicable Information). ACM SIGMOD Record **16** 1 (1987) 42–50.
7. Date, C.J.: Null Values in Database Management. In: Relational Database: Selected Writings. Addisson-Wesley Publishing Company, Reading, USA (1986) 313–334.
8. Date, C.J.: Faults and Defaults. In: Relational Database Writings 1994–1997. Date, C.J., Darwen, H., McGoveran, D. (eds.), Addisson-Wesley Publishing Company, Reading, USA (1998).
9. de Caluwe, R. (ed.): Fuzzy and Uncertain Object-oriented Databases: Concepts and Models. World Scientific, Signapore (1997).
10. de Tré, G.: Extended Possibilistic Truth Values. International Journal of Intelligent Systems **17** (2002) 427–446.
11. Dubois, D., Prade, H.: Possibility Theory. Plenum Press, New York, USA (1988).
12. Dubois, D., Prade H.: Possibility Theory, Probability Theory and Multiple-Valued Logics: A Clarification. Annals of Mathematics and Artificial Intelligence **32** 1–4 (2001) 35–66.
13. Dyreson, C.E.: A Bibliography on Uncertainty Management in Information Systems. In: Uncertainty Management in Information Systems: From Needs to Solutions. Motro, A., Smets P. (eds.), Kluwer Academic Publishers, Boston, USA (1997) 415–458.
14. Klir, G.J., Yuan, B.: Fuzzy Sets and Fuzzy Logic: Theory and Applications. Prentice Hall, New Jersey, USA (1995).
15. Korth, H.F., Silberschatz, A.: Database Research Faces the Information Explosion. Communications of the ACM **40** 2 (1997) 139–142.
16. Bordogna G., Pasi G. (eds.) Recent Issues on Fuzzy Databases. Physica-Verlag, Heidelberg, Germany (2000).
17. Petry, F.E.: Fuzzy Databases: Principles and Applications. Kluwer Academic Publishers, Boston, USA (1996).
18. Prade, H., Testemale, C.: Generalizing Database Relational Algebra for the Treatment of Incomplete or Uncertain Information and Vague Queries. Information Sciences **34** (1984) 115–143.
19. Rescher, N.: Many-Valued Logic. Mc.Graw-Hill, New York, USA (1969).
20. Riedel, H., Scholl M.H.: A Formalization of ODMG Queries. In: Proc. of the 7th Working Conference on Database Semantics (DS-7), Spaccapietra S., Maryanski F. (eds.), Leysin, Switzerland, (1997) 63–90.
21. Zadeh, L.A.: The concept of linguistic variable and its application to approximate reasoning Parts I, II and III. Information Sciences **8** 199–251, **8** 301–357, **9** 43–80 (1975).
22. Zadeh, L.A.: Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets and Systems **1** (1978) 3–28.