data integration, taxonomy of concepts, tuple format,
XML format, medical diagnosis, treatment method,
algebra of medical concepts

Beata M. JANKOWSKA*

MEDICAL DATA INTEGRATION – AN ALGEBRAIC APPROACH

It would be difficult to find a knowledge domain in which data integration is as important as in medicine, being interested in human health and life. Firstly, medical data integration enables acquiring all the information, stored anywhere, about a patient who needs an urgent medical intervention. Secondly, thanks to the integration, we can obtain an aggregate data comprising all known, similar medical cases. In fact, the aggregate data is a source of medical knowledge and a base for formulating conclusions about hypothetical diagnoses, and methods of treating the disorder diagnosed. In the paper, it is shown that a uniform, theoretical approach for the both kinds of data integration: horizontal and vertical, respectively, is possible. It is an algebraic approach, joined together with the use of taxonomy of medical and related concepts. Having the algebra of the concepts, we can as well check the permissibility of data integrating (by means of the subsumption relation, defined on the set of all concepts), as describe the semantics of this process (by means of the usual algebraic operations of sum and intersection).

1. INTRODUCTION

The problem of data integration belongs to those burning issues that are being intensively examined and somehow resolved. It can be considered from various points of view. From one hand - we try to achieve the best possible result of data integration (effectiveness), from the other one - to minimize duration of the whole integration process (efficacy). The goals mentioned above are usually in contradiction of each other. Considering that the amount of data stored in electronic formats is growing fast and fast, and, what is more - these formats are often similar in syntax, the possibility of automatic or semi-automatic data integrating increases. That is why, a high level of data distribution is not any obstacle for their exploring and integrating.

One of the knowledge domains that have to deal with a significant data distribution is medicine. A large amount of medical data, acquired by means of experiments, is stored in numerous hospital repositories and also in, much more numerous, electronic patients' files maintained by doctors' offices. We can easily imagine a case when the data stored in one place turn out insufficient for making the right diagnostic or therapeutic decision: either because of their incompleteness or because of their low reliability. At that time, well done data integration could be a remedy for resolving this problem.

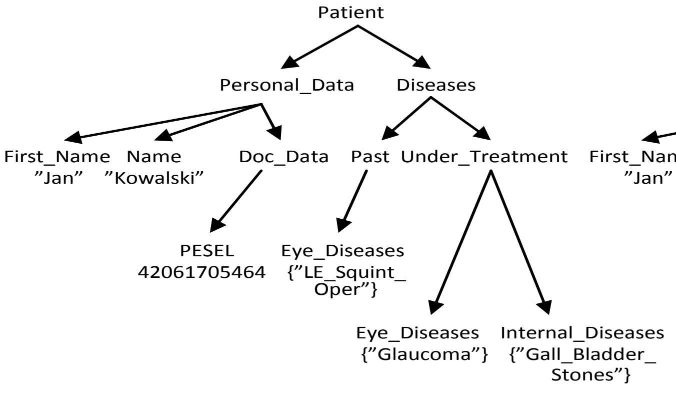
To begin with, let us imagine a situation when an elderly patient, complaining of visual-field disorder, visits an eye doctor's office. After having the patient examined, the eye specialist makes a diagnosis of open-angle glaucoma in him. He thinks that patient's sight troubles are serious enough to administer an aggressive pharmacotherapy (or even to perform laser surgery). The projected pharmacotherapy is, however, absolutely excluded in a case of concurrent sharp cardiac insufficiency. In the past, the patient has undergone treatment because of some heart disease. He has been even hospitalized in a cardiology ward, but he cannot remember the name of his disease. In such the situation, the eye specialist takes up an effort to search out the necessary data in an adequate database, or even - in all the medical databases which he is entitled to access.

For a change, let us consider another case when a specialist in allergology, doctor remaining on hospital duty, admits a child with a fatal asthma exacerbation. He could order a routine clinical treatment, consisting in administering intravenous steroids to the child, if only it was not the case of concurrent strong diabetes. A few months earlier the doctor was met with a very similar case in his practice, and - he remembers this fact well enough - a girl of nine, in the course of such steroid treatment, went into diabetic coma. That is why, the doctor is searching for other similar cases in medical repositories. He wants to verify his hypothesis about the reason for that unfortunate situation.

2. DATA INTEGRATION - HORIZONTAL OR VERTICAL?

Most of medical data are not simple but structured ones. Such the data can be displayed by means of tree structures. The case of the ophthalmologic patient mentioned above consists in widening his initial data record, being in possession of the eye doctor's office (Fig.1), by necessary cardiologic data, being in possession of a hospital cardiology ward (Fig.2). For this data integration to be permissible, it is sufficient and necessary that the personal data of the both patients considered are equal. The result of such an integration would be a data record comprising more information than that initial one (Fig.3). As regards tree structures, the integration would result in (irregular) extension of the initial data tree structure, mainly – in "horizontal" direction. Apart from the global change of the tree structure (qualitative change), there could also appear local changes in tree nodes. We mean here the increase of the set of values attached to an existing leaf node (quantitative change).

* Poznań University of Technology, Institute of Control and Information Engineering, 60-965 Poznan, Pl. M.Sklodowskiej-Curie 5



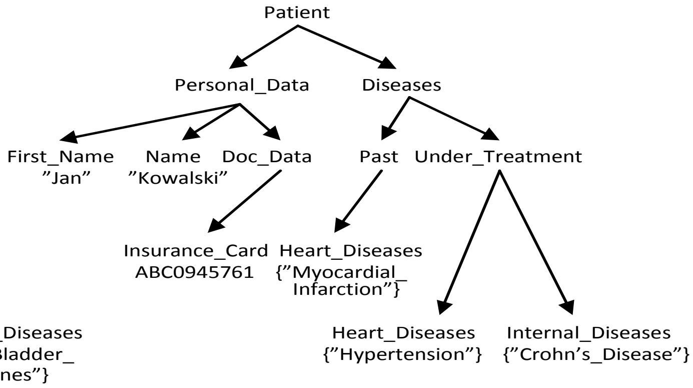

Patient
Personal_Data
Diseases
First_Name ”Jan”
Name ”Kowalski”
Doc_Data
Past
Under_Treatment
Insurance_Card ABC0945761
Heart_Diseases {”Myocardial_ Infarction”}
Heart_Diseases {”Hypertension”}
Internal_Diseases {”Crohn’s_Disease”}

Fig.1. Ophthalmology patient record

Fig.2. Cardiology patient record

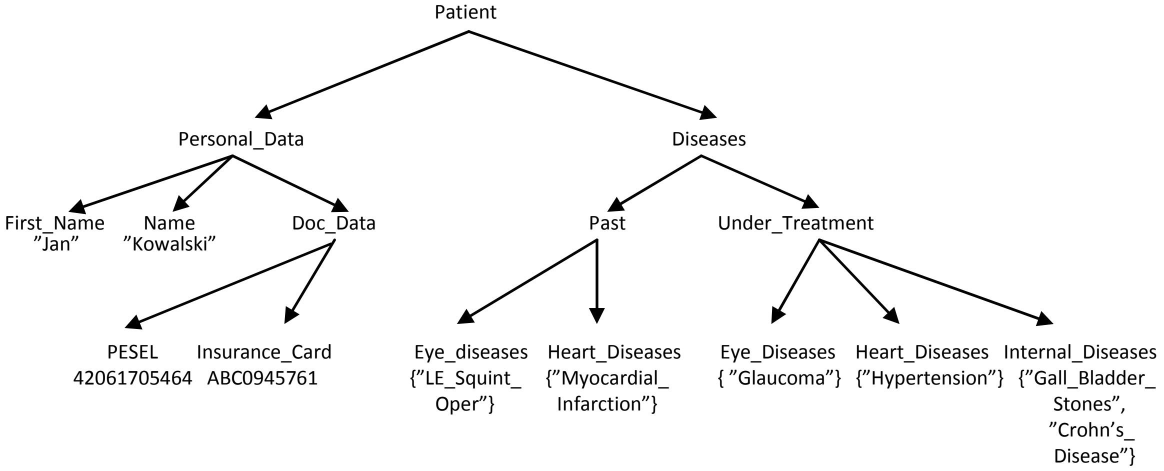

Patient
Personal_Data
Diseases
First_Name ”Jan”
Name ”Kowalski”
Doc_Data
Past
Under_Treatment
PESEL 42061705464
Insurance_Card ABC0945761
Eye_diseases {”LE_Squint_ Oper”}
Heart_Diseases {”Myocardial_ Infarction”}
Eye_Diseases { ”Glaucoma”}
Heart_Diseases {”Hypertension”}
Internal_Diseases {”Gall_Bladder_ Stones”, ”Crohn’s_ Disease”}

Fig.3. Integrated patient record

In turn, the case of the child suffering from the fatal asthma exacerbation requires not completing the child’s data record, but rather searching out similar records, relating to other children. Having compared the records, we should select from among those meeting the requirements of diagnosed (and being under treatment) both bronchial asthma and diabetes. Obviously, the other corresponding records’ attributes can differ to some extent one from another.

As a result of the comparison mentioned above a temporary aggregate record will be obtained, with an attribute stating the number of partial records that have been integrated in this aggregate one: the greater the number, the greater the aggregate record’s reliability. The values of its specific attributes can be calculated by intersecting appropriate values over all partial data records. For the sake of its nature, such a data integration can be called as ”vertical” or ”temporary”, in contrast to the previous one, called ”horizontal” (unlike at Anjum et al. [1]!). It is worth noticing, that the aggregate record is a kind of data being made to order, and its state depends on the current state of databases which partial data have been acquired from. Such the aggregate record cannot be subjected to any update; at the most, it can be replaced by quite a new aggregate record.

3. XML FORMAT OF MEDICAL DATA REPRESENTATION

The widely understood data exchange and integration can be effective if and only if all the data considered are represented in electronic formats that are equal or similar to one another. This obvious observation has been a base for defining standards of domain knowledge representations. Among various important domains of knowledge, also medial knowledge has been standardized. The first proposal of such a standard [6] was put forward in 1987, by the Health Level Seven (HL7) organization. Since the late nineties, it has been developed under the common banner of Electronic Health Records (EHRs) [4]. After all, it has not become general yet. Instead of this, many local standards, adjusted to national health care systems (e.g. Polish NHF Standard [9]), or detailed standards, related to specialized medical fields (e.g. SNOMED CT [12], Disease Ontology [3]) have been introduced. What is important, most of them work with the technology of XML [5].

The XML data are being constructed according to meta-rules, called ”syntactic schemas”, given for setting the detailed relationships between data components. These relationships can differ (slightly or even considerably) between different formats of knowledge representation. If integration can be performed also in respect to data of different formats, it is obvious that they are not these relationships which determine the permissibility and the course of integrating. In order to express what are data integration constraints, we can use data attributes themselves. This observation lead us to believe that multi-level XML data can be flattened – without a loss in their semantics – to the form of one-level data tuples. This form will be more convenient while establishing requirements for data integration and, next, while implementing the process of data integration. In a case of horizontal data integration, the obtained aggregate tuple would have to be inversely transformed to the structured XML form. However, in a case vertical integration, such an inverse operation would be redundant.

Let us reconsider the data from the Figures 1 and 2. The above data can be transformed to the form of the following two tuples (1) and (2).

(Participants=0; First_Name=”Jan”; Name=”Kowalski”; PESEL=42061705464;
P_Eye-Dis={”LE_Squint_Oper”}; UT_Eye_Dis= {”Glaucoma”}; UT_Int_Dis={”Gall_Bladder_Stones”}) (1)

(Participants=0; First_Name=”Jan”; Name=”Kowalski”; PESEL=42061705464; Ins_Card= ABC0945761;
P_Heart_Dis= {”Myoc_Infarction”}; UT_Heart_Dis= {”Hypertension”}; UT_Int_Dis={”Crohn’s_Disease”}) (2)

The *Participants*, listed at the first place among tuple’s attributes, will be commented later on. The tuple (3), obtained as a result of integration (1) and (2), can be easily transformed to the XML data shown in the Figure 3, but on the condition of knowing this data scheme.

(Participants=0; First_Name=”Jan”; Name=”Kowalski”; PESEL=42061705464; Ins_Card= ABC0945761;
P_Eye-Dis={”LE_Squint_Oper”};P_Heart_Dis={”Myoc_Infarction”};UT_Eye_Dis={”Glaucoma”};
UT_Heart_Dis={”Hypertension”}; UT_Int_Dis={”Crohn’s_Disease”, ”Gall_Bladder_Stones”}) (3)

In order to illustrate the course of vertical integration, let us imagine an evidence of some clinical research that was done on two subgroups (A) and four subgroups (B) of young asthmatic patients, with concurrent moderate or severe diabetes, and fatal asthma exacerbation diagnosed on the spot. Let us assume that the evidence has a form of register of clinical trials from the Cochrane Library. The patients mentioned were administered a new drug, called as ”new GCS”, of unknown efficiency (a treatment group) or a traditional one, called as ”traditional GCS” (a control group). Two clinically essential outcomes were tested during the experiment: the length of the current hospital stay (not longer or longer than 7 days) and the necessity of next hospital admission (before or upon half a year). The obtained results have been put together in a usual summary table (Tab. 4).

Table 4. A medical experiment on asthmatic children - summary results

Comparison: New GCS (multiple doses) + Beta-2-Agonist vs Traditional GKS + Beta-2-Agonist
Outcome: Current hospital stay above 7 days (left); Next hospital admission before half a year (**right**)
Co-intervetion: Antidiabetes drugs

Study	**Treatment Group** $\hat{p}_1 = \frac{A_1}{N_1}$	**Control Group** $\hat{p}_0 = \frac{A_0}{N_0}$	**RR 95%CI RR 95%CI**
ExperimentA, Group A1	4/17 **3/17**	6/14 **7/14**	0.55 (0.19, 1.57) **0.35 (0.11, 1.12)**
ExperimentA, Group A2	5/12 **2/12**	6/13 **4/13**	0.90 (0.37, 2.20) **0.54 (0.12, 2.44)**
ExperimentB, Group B1	11/52 **5/52**	18/49 **11/49**	0.58 (0.30,1.09) **0.43 (0.16, 1.14)**
ExperimentB, Group B2	7/35 **8/35**	7/36 **14/36**	1.03 (0.40, 2.63) **0.59 (0.28, 1.22)**
ExperimentB, Group B3	9/28 **7/28**	12/24 **10/24**	0.64 (0.33, 1.26) **0.60 (0.27, 1.33)**
ExperimentB, Group B4	6/41 **4/41**	10/36 **11/36**	0.53 (0.21, 1.31) **0.32 (0.11, 0.92)**
Total	42/185 **29/185**	59/172 **57/172**	0.66 (0.47, 0.93) **0.47 (0.32, 0.90)**

At the same time, the results have been written in a form of structured data, by means of an XML format specific for group experiments. For example, those relating to the first group (A) of the experiment participants will be expressed as follows (4):

```
<experiment_evidence>
  <treatment_group>
    <participants>
      <global_number>29</global_number>
      <age_range  group=”children” min=”7” max=”18”></age_range>
      <others race = ”white black”></others>
      <main_disease>asthma</main_disease>
      <concurrent_diseases important=”diabetes”></concurrent_diseases>
    </participants>
```

```
<pharmacological_trethement>
  <basic_drug_used product=”new_GCS” min_daily_dose=”160μg” max_daily_dose=”320μg”></basic_drug_used>
  <other_drugs_used others1=”Beta-2-Agonist”></other_drugs_used>
</pharmacological_treatment>
<experiment_results>
  <hospital_stay_length>
    <result_p min=”4-days”  max=”6-days” number=”20”></result_p>
    <result_p min=”8-days”  max=”23-days” number=”9”></result_p>
  </hospital_stay_length>
  <necessity_of_next_hospitalization>
    <result_p min=”3-months”  max=”6-months” number=”5”></result_p>
    <result_p min=”8-months” max=”31-months” number=”24”></result_p>
  </necessity_of_next_hospitalization>
</experiment_results>
</treatment_group>
.....
</experiment_evidence>
```

(4)

After having transformed the above XML data and the analogous one, including the results of the second group (B) of the experiment’s participants, to the form of one-level data tuples, we obtain (5) and (6), respectively.

(Global_Number=29; Main_Dis={”Bronch_Asthma”}; Conc_Dis={”Diabetes”}; Age_Category={”Children”};
Basic_Drug_Used={”New_GCS”}; Alter_Treat ={”Traditional_GKS”}; Comp_Treat={”Beta-2-Agonist”};
Drug_Dosage=<160 ; 320>; Age_Range=<7; 18>; Race={”White”, ”Black”}; Part_With_Hsl+=9; Part_With_Nnh-=5)(5)

(Global_Number=156; Main_Dis={”Sev_Bronch_Asthma”}; Conc_Dis={”Diabetes”}; Age_Category={”Children”};
Basic_Drug_Used={”New_GCS”}; Alter_Treat={”Traditional_GKS”}; Comp_Treat={”Beta-2-Agonist”};
Drug_Dosage=<160 ; 400>; Age_Range=<5; 17>; Part_With_Hsl+=33; Part_With_Nnh-=24) (6)

Next, we find it permitted to perform the integration of the tuples (5) and (6) (”Sev_Bronch_Asthma”, being the value of the attribute *Main_Dis* from (6) is a specific case of the more general ”Bronch_Asthma”, being the value of the same attribute from (5), and the values of remaining important attributes, i.e. *Conc_Dis*, *Age_Category*, *Basic_Drug_Used*, *Alter_Treat*, *Comp_Treat* are equal in the both tuples), and, after having done it, we achieve an aggregate tuple (7).

(Global_Number=185; Main_Dis={”Bronch_Asthma”}; Conc_Dis={”Diabetes”}; Age_Category={”Children”};
Basic_Drug_Used={”New_GCS”}; Alter_Treat ={”Traditional_GKS”}; Comp_Treat={”Beta-2-Agonist”};
Drug_Dosage=<160 ; 400>; Age_Range=<5; 18>; Part_With_Hsl+=42; Part_With_Nnh-=29) (7)

The global number of the children subjected to the experiment was high enough to find the final result reliable and to formulate a hypothesis (8) about treating a child, suffering from asthma and diabetes, in a case of fatal asthma exacerbation.

```
asthma_treatment_rule        g-PM 0.84:
   if   fatal_asthma_exacerbation and
            strong_diabetes and
      age_range [5 ; 18] and
      basic_drug new_GCS[160 ; 400] instead_of traditional_GCS and
      complementary Beta-2-Agonist
   then hospital_stay_length≥7            PM-c 0.68
        next_hospitalization≤6   PM-c 0.49
```

(8)

This favourable situation is an exception, not a rule. In most cases, in order to ”gather” enough experiment participants, it is necessary to search out and integrate a great number of medical data, also those of individuals, being treated at doctors’ offices.

The hypothesis can be inserted (temporarily or permanently) into the knowledge base of an expert system supporting doctors, specialists in allergology, in making their decisions [8]. Then, all the tuples obtained, also – aggregate ones, become needless and should be removed from the working memory.

4. AN ALGEBRA FOR STUDYING REQUIREMENTS AND THE PROCESS OF DATA INTEGRATION

If medical data are stored in various electronic formats, then the permisibility of their integrating can be examined only when knowing the schemes of these data. Next, in order to perform the integration, it is necessary to define formal mappings between the schemes and instances. As regards this meta-knowledge, essential for carrying out the process of integration, it is now the subject of intensive scientific research (e.g. [11, 10]). However, it is only a slight extent to which we make use of ontology knowledge that could provide the possibility of flexible defining of both constraints and results of data integration process [2]. It is probable that, taking advantage of ontology, we would be able to integrate some of the data which are now considered as nonintegrable. The integrated data would be given the factors of reliability, depending on the similarities between data components. Let us notice that, at the lack of any serious medical hypothesis, even a medium reliable null hypothesis is a good starting point for further research.

In order to increase the flexibility of data integration, we suggest defining an algebra of medical (restricted to a chosen medical specialty) and related concepts [7]. The algebra can be established based on the concepts' taxonomy. Having the algebra, we will check the permissibility of data integrating – by means of the subsumption relation between the concepts, and we will define the semantics of this process – by means of algebraic operations.

Let CS stand for the set of all such concepts. Some of them are "abstract" ones (e.g. Eye_Disease, Antiglaucoma_Drug), while the remaining are "real" ones (e.g. Adults_Range, Prostaglandin). Between any pair of abstracts concepts and any pair of real concepts the relation "is_more_general_than" can hold (e.g. Eye_Disease is_more_general_than Genetic_Eye_Disease and Adults_Range is_more_general_than Elderly_People_Range). In turn, between a real concept and an abstract concept the relation "is_an_instance_for" can hold (e.g. Prostaglandin is_an_instance_for Antiglaucoma_Drug). The real concepts represent individuals, being numbers or strings, or sets of such individuals (e.g. the concept Name takes its values from the finite set of strings like "Kowalski", "Nowak" or "Jankowska"). These last concepts have to be considered together with the sets representing their values (e.g. Name{"Kowalski", "Nowak", "Jankowska", etc}). The interpretation of this attached set of values depends on a context of its use: it will denote either a set of features of the concept (the greater number of set elements, the more expressive power of the concept), or a set of alternative forms of the concept (the greater number of set elements, the less expressive power of the concept).

Actually, classifying the concept as belonging to abstract or real ones is a matter of discretion. However, it has deep consequences on the form of structured data: at their lowest levels (of "leaves") only real concepts can occur.

Let us denote by I the set of all the individuals mentioned above. The subsumption relation $\leq_{CS}$, specified on the domain set CS, is a partial order relation (corresponding to the informal relations is_more_general_than and is_an_instance_for), complying with the following requirement (9):

$$\forall (rc \in CS)\ \forall (i_{11}, i_{12}, \ldots, i_{1m}, i_{21}, i_{22}, \ldots, i_{2n} \in I)$$

$$((rc\ \{i_{11}, i_{12}, \ldots, i_{1m}\} \leq_{CS} rc\ \{i_{21}, i_{22}, \ldots, i_{2n}\}) \leftrightarrow ((\{i_{21}, i_{22}, \ldots, i_{2n}\} \subseteq \{i_{11}, i_{12}, \ldots, i_{1m}\}) \vee (\{i_{11}, i_{12}, \ldots, i_{1m}\} \subseteq \{i_{21}, i_{22}, \ldots, i_{2n}\}))) \quad (9)$$

where *rc* stands for a real concept with an attached set of individuals.

Let us add into the set CS two additional concepts: the most general one, denoted by $\top$, and the most specific one, denoted by $\bot$. The obtained domain set CS can be modelled by means of a weakly connected directed graph, in which nodes represent concepts and directed edges – the relation $\leq_{CS}$. Let us assume that, on the set CS, the operations of sum $\cup_{CS}$ and intersection $\cap_{CS}$ are defined. They are as follows:

$$\forall (c_1, c_2 \in CS)\ (c_1 \cup_{CS} c_2 = sup\{c_1, c_2\}) \quad (10)$$

$$\forall (c_1, c_2 \in CS)\ (c_1 \cap_{CS} c_2 = inf\{c_1, c_2\}) \quad (11)$$

Obviously, the algebra **CS** = (CS, $\cup_{CS}$, $\cap_{CS}$) is a lattice. This way, we obtain a strong tool to compare concepts from the domain CS and to perform algebraic operations on them, and, as a consequence, to check the constraints of the integration process and to calculate its results.

In order to illustrate our considerations, let us deal with medical data being in the form of specific tuples. Namely, each tuple Tup_i has a three-segment structure, of the form (12):

$$Tup_i = <a_i=n_i;\ b_{i1}=v_{i11};\ b_{i2}=v_{i12};\ \ldots;\ b_{im}=v_{i1m};\ c_{i1}=v_{i21};\ c_{i2}=v_{i22};\ \ldots;\ c_{in}=v_{i2n}>, \quad (12)$$

where a_i stands for an attribute denoting the number (n_i) of partial data, having been integrated into the tuple (a_i=0 stands for a tuple that can be neither a subject nor a result of vertical integration), b_{ik} and c_{il} – for attributes of primary and secondary importance in the considered tuple, respectively, and v_{i1k}, v_{i2l} – for real concepts, being values of b_{ik} and c_{il}, respectively.

Most often, we consider the following two tuples Tup_{i1} and Tup_{i2}:

$$Tup_{i1} = <a_i=n_{i1};\ b_{i1}=v_{i11};\ b_{i2}=v_{i12};\ \ldots;\ b_{im}=v_{i1m};\ c_{i1}=v_{i1(m+1)};\ c_{i2}=v_{i1(m+2)};\ \ldots;\ c_{in}=v_{i1(m+n)}>$$

$$Tup_{i2} = <a_i=n_{i2};\ b_{i1}=v_{i21};\ b_{i2}=v_{i22};\ \ldots;\ b_{im}=v_{i2m};\ c_{i1}=v_{i2(m+1)};\ c_{i2}=v_{i2(m+2)};\ \ldots;\ c_{in}=v_{i2(m+n)}> \quad (13)$$

to be integrable if and only if the constraint (14) is satisfied

$$(n_i = 0 \leftrightarrow n_j = 0) \wedge \forall (1 \leq k \leq m)\ (v_{i1k} \leq_{CS} v_{i2k}) \quad (14)$$

The constraint imposed on component tuples excludes the possibility of integrating ”individual” tuples (of $a_i = 0$) with ”aggregate” tuples (of $a_i > 0$), and reversely. Besides, the restrictions regarding the attributes of primary importance have to be satisfied: for each attribute b_{ik}, its value v_{i1k} in the first (initial) tuple has to be more general than its counterpart v_{i2k} in the second tuple, being joined onto the initial one.

Obviously, the constraints imposed on the tuple integration process can be completely reformulated, admitting to treat all the tuple's parameters differently, each one in an appropriate way. In particular, we can demand the values v_{i1k} and v_{i2k} of the attribute b_{ik} be equal $((v_{i1k} \leq_{CS} v_{i2k}) \wedge (v_{i2k} \leq_{CS} v_{i1k}))$.

In order to define tuple integration results, let us make use of the operations from the algebra CS. Most often, the result of integration of the two tuples Tup_{i1} and Tup_{i2} (13) will be defined as a tuple Tup_{i3} (15), in a case of $n_{i1}=n_{i2}=0$, and as a tuple Tup_{i4} (16), in a case of $n_{i1} \neq 0$, $n_{i2} \neq 0$:

$$Tup_{i3} = <a_i=0;\ b_{i1}=v_{i11} \cup_{CS} v_{i21};\ b_{i2}=v_{i12} \cup_{CS} v_{i22};\ \ldots;\ b_{im}=v_{i1m} \cup_{CS} v_{i2m};$$

$$c_{i1}=v_{i1(m+1)} \cup_{CS} v_{i2(m+1)};\ c_{i2}=v_{i1(m+2)} \cup_{CS} v_{i2(m+2)};\ \ldots;\ c_{in}=v_{i1(m+n)} \cup_{CS} v_{i2(m+n)}> \tag{15}$$

$$Tup_{i4} = <a_i=n_{i1}+n_{i2};\ b_{i1}=v_{i11} \cap_{CS} v_{i21};\ b_{i2}=v_{i12} \cap_{CS} v_{i22};\ \ldots;\ b_{im}=v_{i1m} \cap_{CS} v_{i2m};$$

$$c_{i1}=v_{i1(m+1)} \cap_{CS} v_{i2(m+1)};\ c_{i2}=v_{i1(m+2)} \cap_{CS} v_{i2(m+2)};\ \ldots;\ c_{in}=v_{i1(m+n)} \cap_{CS} v_{i2(m+n)}> \tag{16}$$

In the light of the above proposal, let us reconsider the integration of the exemplary tuples (1) and (2) from the chapter 3. At the beginning, let us divide the set of tuple attributes into three disjoint subsets: one-element subset {*Participants*}, containing the attribute of the sort a_i, a subset including attributes that are of primary importance {*First_Name*, *Name*, *PESEL*}, and a subset including attributes that are of secondary importance {*Insurance_Card*, *P_Eye_Disease*, *UT_Eye_Disease*, *P_Heart_Disease*, *UT_Heart_Disease*, *UT_Int_Disease*}. Since the attribute *Participants* takes a value of zero in the both tuples, then only horizontal integration is thinkable.

Let us notice the absence of *Insurance_Card*, *P_Heart_Disease* and *UT_Heart_Disease* among the attributes listed in the tuple (1), and also the absence of *P_Eye_Disease* and *UT_Eye_Disease* among the attributes listed in the tuple (2). It means that all these attributes take a value of $\top$.

The integration of the tuple (2) with the tuple (1), that was initially recognised as intuitively possible, can be now checked formally. After having found out that the values of the corresponding attributes of primary importance are equal in the both tuples, we come to the conclusion that the constraint (14) is fully satisfied, and – as a consequence – the tuple (2) is integrable with the tuple (1). Next, according to the definition proposed (15), the process of integration will consist in summing the values of the corresponding attributes (of both primary and secondary importance) by means of the algebraic operator $\cup_{CS}$. Remembering that:

- for any (real or abstract) concept c from CS, it holds: $c \cup_{CS} \top = \top \cup_{CS} c = c$, and
- in an ”individual” tuple (of a_i=0), the less set of values, the more general concept with this set attached, as a final result we obtain a tuple equal to that expected one (3).

5. SUMMARY

The proposed algebraic approach to the problem of medical data integration has such an advantage that the integration can be interpreted more extensively than up to now. According to this approach, the term ”integration” covers as well ”complementing” data by missing partial values, as ”joining” similar data, in order to increase their expressive power. The rules of integration, comprising both checking the constraints of the integration process and calculating its results, are flexible. The implementation of these rules is based on the knowledge of data semantics, which is expressed by means of medical concepts taxonomy, and the use of operators of the algebra proposed.

Before having started integration by means of the algebraic method, one should build a taxonomy of medical and related concepts for a chosen medical specialty. Particularly carefully, the classification of concepts into the groups of abstract and real ones should be performed. Also, it is worth noticing, that the process of integration has to be preceded by the initial preprocessing of data, which are being stored in different electronic formats and embedded in different, although similar, ontologies. In order to obtain the form of tuples required in the algebra domain, it is necessary to transform the data (ontology- and taxonomy-driven transformation) to just this form.

ACKNOWLEDGEMENT

The research has been supported in part by The Polish Ministry of Science and Higher Education under Grant N516 015 31/1553

BIBLIOGRAPHY

[1] ANJUM A. et al., The Requirements for Ontologies in Medical Data Integration: A Case Study, in: Proc. 11th Int. Database Engineering and Applications Symposium (IDEAS 2007), pp.308-314, IEEE Computer Society, Canada, 2007.

[2] BERLANGA R., Medical Data Integration and the Semantic Annotation of Medical Protocols, in: Proc. 21st IEEE Int. Symposium on Computer-Based Medical Systems, pp.644-649, IEEE Computer Society, 2008.

[3] Disease Ontology, http://diseaseontology.sourceforge.net/, 2007.

[4] EHR Standards, http://www.openehr.org/standards/iso.html, 2005.

[5] Extensible Markup Language (XML), http://www.w3.org/XML/, 2004.

[6] HL7 Standard, http://www.interfaceware.com/manual/hl7.html, 2007.

[7] JANKOWSKA B.M., Specificity and Methods of Medical Data Integration, Polish Journal of Environmental Studies, vol.17, no 2A, pp. 24-28, Olsztyn, 2008.

[8] JANKOWSKA B., SZYMKOWIAK M., How to Acquire and Structuralize Knowledge for Medical Rule-Based Systems?, in: Studies in Computational Intelligence, vol.102, pp.99-116, Springer Berlin/Heidelberg, 2008.

[9] Otwarty standard wymiany komunikatów NFZ (NHF Open Standard of Message Exchange), http://www.nfz.gov.pl/new/index.php?katnr=0&dzialnr=15&artnr=2347, 2007.

[10] PANKOWSKI T., XML Data Integration in SIXP2P: A Theoretical Framework, in: ACM International Conference Series, vol.261, pp.11-18, ACM, NY, 2008.

[11] ROTH M. et al., XML mapping technology: Making connections in an XML-centric world, in: IBM Systems Journal, vol.45, no 2, pp.389-410, 2006.

[12] SNOMED Clinical Terms. Users Guide, http://www.ihtsdo.org/fileadmin/user_upload/Docs_01/Technical_Docs/snomed_ct_user_guide.pdf, 2007.