Ontology Based Statistical Automated Inference - New Approach to Artificial Intelligence

Wlodzimierz Borkowski* and Hanna Mielniczuk

Independent Research Group Warsaw, Poland

Abstract: Statistical analysis requires understanding the nature of the phenomenon under study, as well as understanding sense of mathematical statistics. Bridging the gap between semantic web based on knowledge representation languages, and concepts described by mathematical formula is a challenge for AI. In order to overcome this gap the ontology language P-ONT (based on directed graph) has been invented. To illustrate the capabilities of the P-ONT language, semantic web (built on the P-ONT ontology) OLAP cube, relational data bases and generalized hierarchical statistical regression models are presented.

Keywords: Ontology, AI, OLAP Cube, General Linear Model, Statistical Inference, Hierarchical Statistical Model.

1. INTRODUCTION

Today in epidemiology takes place evolution that leads from an individual epidemiological study to generalized understanding phenomenon of health (in particular any disease phenomenon). There is observed significant growth in the understanding and synthesis of epidemiological concepts. A systematized body of principles by which to design and evaluate epidemiology studies began to form in the second half of the 20th century [1]. Nevertheless, it is believed that epidemiology remains in early stage of development [2]. Analyzing probabilistic regularities in experimental and observational epidemiology is carried out using statistical techniques and subjects them to the concepts of causation.

Statistical analysis is developed independently of epidemiology and focused on increasingly sophisticated techniques. Furthermore statistics brings together specific technique (and hidden behind their statistical models) with the organization of study providing probabilistic interpretation of the results of the statistical analysis. Statistics has its origin in the work of the R. Fisher in experimental agriculture studies and still the experiment is the mainstream of statistics. In medicine is the same. Despite the fact that clinical randomized trials allow assess only efficacy, they have a much higher level of evidence [3] than observational studies allowing an evaluation of an effectiveness. The development of medical informatics and statistics give hope that the observational studies may have a high level of evidence. A high level of evidence of a clinical trial results not only from randomized sampling but also from the formalized description of individual patients and strict adherence to sampling design. Currently there are formed large collections of clinical data encoded according to the UMLS (Unified Medical Language System) [4] and by strict organizational rules (for example, in Quality Assurance or IV phase clinical trial). Level of evidence of observational studies, where the rational complex sampling was used may not be lower than randomized trials. While experimental studies require first of all to determine the correct design and discipline during their execution, in observational studies there is a need of understanding the evaluated health phenomenon, as well as the sense of advances statistical technique (for inference and sampling) and causal reasoning.

Issues outlined above constitute an inspiration for us to work on a statistical automated inference in the field of AI.

Our approach is alternative to Data Mining. Data Mining discovers previously unknown patterns from large data sets, whereas statistical inference draws conclusions from data sets subjected to random variations. The essential difference among them lies in the fact that data mining focuses on pattern discovering, while statistical inference reveals the patterns and moreover shows the rules determining random events subordinated to the patterns.

In our opinion, data mining, applying classification and clustering of different provenance (also statistical factorial analysis and logistic regression), suffers from the lack of understanding of health phenomenon and the lack of causation.

Automated inference proposed by us contains various aspects. The first is a deductive inference generally based on descriptive languages (family of

*Address corresponding to this author at the Independent Research Group Warsaw, Szczygla 13, 05-420 Jozefow Poland; Tel: 48-506612435;
E-mail: wlodzimierz.borkowski@gmail.com

E-ISSN: 1929-6029/12

first order languages) [5-7]. The second refers to an experimental study in the real world recognized by mathematical concepts. Mathematical statistics rests on two cornerstones. The first is a linear algebra (matrix calculus), the second is a theory of probability in the space R^n. It should be noted that both the ontology of linear algebra and the ontology of probability are not yet researched, as far as we know. Bridging the gap between semantic web based on knowledge representation languages, and the mathematical concepts described by mathematical formula is a challenge for artificial intelligence. In order to overcome this P-ONT language has been invented by us. P-ONT is generic first order language (metalanguage for it is directed graph) [8, 9]. Next aspect of a statistical analysis is data management of the OLAP Cubes and relational data bases. Controversial probabilistic ontology is discussed and our understanding in this regard is shown. The knowledge base about considered health phenomenon is a concept of great importance.

The knowledge base contain a set of causal quantitative relations recorded by real numbers. As a backbone the knowledge base we have used hierarchical generalized linear model (described as P-ONT semantic web).

The article presents our accomplishment concerning these issues. Standard ANOVA with interaction and Linear Regression as well as General and Generalized Linear Model are considered. Then the a hierarchical models are presented. The course of the automated statistical inference is shorty presented including: choice of a statistical model adequate to observations and mutually to health phenomenon (outtretched in heath knowledge base); sampling from medical observations data; mapping into P-ONT observational data collected in a relational database; fitting chosen statistical model to sampled observations. Algorithm of the OLAP cube generation, as an example of the computer implementation of a P-ONT semantic web, is discussed.

2.THEORETICAL FOUNDATIONS OF P-ONT LANGUAGE

2.1. Building of the P-ONT Language

Let's start with the structure $S = \{U; \rightarrow; \cup, \cap, \backslash\}$

where universe U is a set of elements (in the sense of set theory).

Link "$\rightarrow$" is one-ary assignment $U \rightarrow U$.

Link "$\rightarrow$" as the concept from metalanguage means two-ary relation in the language of first order logic $\varrho\,(*, *) \subset U * U$ [3].

Link "$\rightarrow$" of y to x corresponds to the arrow in the Graph theory or λ function in λ-calculus. Thus S in the language of first order logic is $S = \{U; \varrho\,(*,*); \wedge; \vee; \neg\}$.

We can denote $\sigma\,(*; *; *)$ as: $\sigma\,(z; x; y) \equiv \{z \rightarrow x, z \rightarrow y, x \rightarrow y\}$ $\sigma\,(z; x, y)$ forms $\varrho\,(z, x) \wedge \varrho(z, y) \wedge \varrho\,(x, y)$

In comparison to OWL/RDF [10, 11] we take a different concept of a property. For example, in OWL/ RDF a woman as person's property is written by link labeled by gender between instance person and instance woman. In contrast, we introduce property gender as a "two component" instance of gender with two levels: women and men (formally instance gender has two properties: woman and man) (Figure **1**).

Now we create mutually disjoint three parts of the $U = \{T \cup C \cup I\}$.

T constant element (named Thing) such that: $T \notin (C \cup I)$.

C set of elements named classes (for constants uppercase A, B, C.. ; for variables lowercase a,b,c...).

Set of classes is: $C = \{c : \varrho\,(T, c)\}$

I set of elements named instances (for constants X, Y, Z ..), (for variables x, y, z...).

$I = U \backslash \{C \cup I\}$

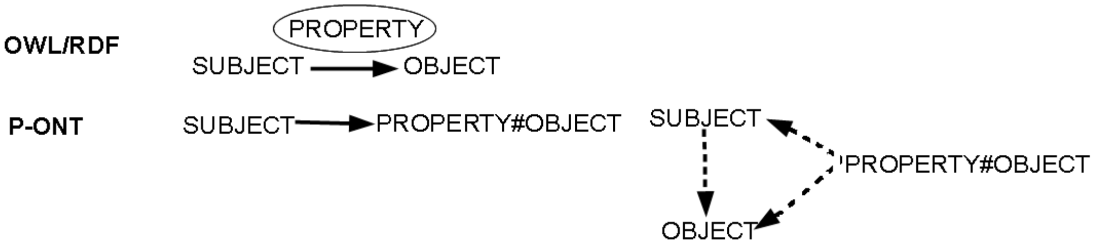


Figure 1: Comparison of the property concept between OWL/RDF and P-ONT.

Definition of the two-ary relation *isa* $(*, *)$.

$isa(*, *) \subset (C * I)$ such that the
$isa(c,x) = \forall x \in I\ \overline{\exists}\ c \in C \varrho(c,x)$

Instances are unique, however, it is convenient to record the class to which the instances are. Thus we have relation between instances (property) $pr\ (*, *) \subset (C * I) * (C * I)$, $pr\ (c, x : d, y) \equiv \forall c \in C\ \overline{\exists} d \in C\ c \neq d\ (\ \overline{\exists}\ x\ isa\ (c\ , x)\ \ \overline{\exists}\ y\ isa\ (d, y)\ \varrho\ (x, y))$. Link $x \rightarrow x$ is forbidden so $pr\ (a, x :a, x)$ is forbidden too.

Definition of the property between classes $prC\ (*, *) \subset (C * C)$:

$prC\ (a, b) \equiv \exists x\ isa\ (a, x)\ \exists\ y\ isa\ (b, y)\ pr\ (x, y)$

$prC\ (a, a) \equiv \exists x, y\ isa\ (a, x) \wedge isa\ (a, y) \wedge \varrho\ (x, y)$

Universe *U* contains two subsets: class and instances plus T as single element. By contrast OWL/RDF has three subsets: classes, instances, properties. Any element of the structure S, we call an ontology (it contains only symbols).

On the basis of a given ontology we built semantic web by attaching real numbers to some instances of the ontology (by "data property"). Set of semantic webs creates universe of the legacy structure. Terms of the legacy structure are equivalent to computer procedures. For example, a single symbol in the ontology (in mathematical formula a parameter) corresponds to a real number, whereas mathematical matrix ontology corresponds to the set of real numbers (acquiring a configuration in accordance with the ontology) with matrix calculus operations.

2.2. Concept of P-ONT Indexing

The concept of P-ONT indexing (Figure **2**) is shown on the two-dimensional contingency table ontology where instances of the class A means the dimensions of contingency table (a1-blood pressure, a2-tobacco smoking), instances of a class B means of the levels of the dimensions (b4-yes, b5-no, b1-high, b2-normal b3-low). Class C contains "two component " instances (c1-pressure; high, c2-pressure; normal, c3-pressure; low, c4- smoking; yes, c5-smoking-no). Class D contains "two component of the two component " instances meaning contingency table indexes. Two dimensional contingency table is written as instance e1 in E class.

Now we want to save the mean value and variance of the distribution of the growth of people with hypertension and cigarette smoking (instance d1 D Class).We save these values in a one-dimensional vector (e2 instance in the class E) where instances d7, d8 means indexes of the fields one-dimensional vector.

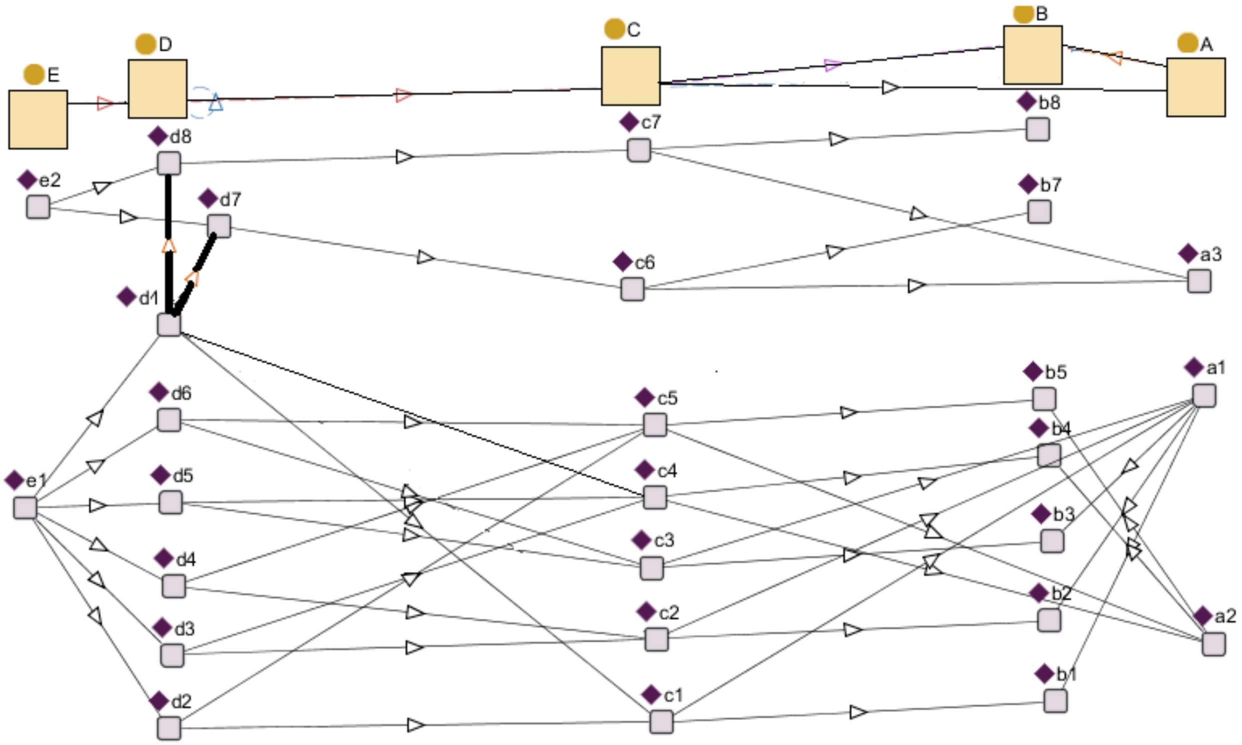


Figure 2: Ontology of the concept a nested indexing (indexing of a index).

2.3. P-ONT Satisfiability

Reasoning as an action in ontology management is an ambiguous concept. We are focused on finding elements of the Semantic Web structure satisfying given P-ONT formula. Our fundamental tool is SPARQL (Query Language for RDF) dedicated originally to descriptive RDF [12]. We have adopted SPARQL to needs P-ONT. Satisfiability (j=) is obtained by queries across instances of the classes. Range of search variables is limited by the defining of the classes and through commands of the FILTER. Hitherto, a RDF triple ontologies were inadequate to the concept described by mathematical formulas [13-15], which cause problems in the automated preparation of a query. SPARQL adapted to the p-ONT allows automated preparing and conducting P-ONT semantic web queries even with high complexity [16].

3. P-ONT SEMANTIC WEB OF OLAP CUBE

OLAP cube is a set of data, organized in a way that facilitates non-predetermined queries for aggregated information (online analytical processing). The OLAP cube consists of numeric facts called measures which are categorized by sets of dimensions.

The different models of OLAP cubes are presented in literature [17-19]. In the partition model linking cubes is the method of overcoming sparsity (not every cell in the cube is filled with data). Instead of creating a sparse cube, it is sometimes better to create another separate, but linked, cube in which a sub-set of the data can be analyzed into great detail. The linking ensures that the data in the cubes remain consistent. The dimensions of the OLAP cube can be organized as a hierarchy and could be summarized using it. Common operations on OLAP cubes include slice and dice, drill down, roll up. Drilling down/up is a specific technique whereby the user navigates among levels of data ranging from the most summarized (up) to the most detailed (down).

Consider the OLAP cubes Cube_C1, Cube_ C2 with two dimensions a1, a2 (Figure **3**). Dimension a1 is hierarchical with two levels: on the first level b1, b2, on the second b1 is splitted into b3, b4, and b2 into b5, b6. Dimension a2 is one-level and has values of b7, b8. Drilling down Cube_C1 of coarser granularity leads to the Cube_C2 of finer granularity.

We use the P-ONT language for modeling the OLAP cubes and perform OLAP operations in this model. Leaving in OLAP_Ontology only instances of class D with linked (by data property) real numbers avoids the situation of sparse cubes.

Mentioned above OLAP cubes are saved in the P-ONT OLAP_Ontology (Figure **4** top) as instances of class F (cubes): Cube_C1 as instance f301 and Cube_C2 as instance f302. The dimensions of the Cube_C1 and Cube_C2 are saved as instances a1, a2 of class A. The OLAP measure would be realized in the OLAP semantic web as data property of class E with range in R, is not a part of the OLAP_Ontology.

Linked to f301 are instances b1, b2 of class B, linked to the f302 instances b3, ..., b6. The instances of class F have no order. In P-ONT OLAP_Ontology drilling is the navigation between instances of the class F.

We introduced also the Dimension Ontology in P-ONT Language (Figure **4** bottom). The nesting of the cubes in the Dimension_Ontology reflect hierarchy of dimensions of the OLAP_Ontology. In the Dimension_Ontology is saved an order of the OLAP_Ontology cubes allowing to perform basic operations of OLAP. Note that the instances a1, a2 being the instances of the class A in OLAP_Ontology are at the same time instances of the class B in the

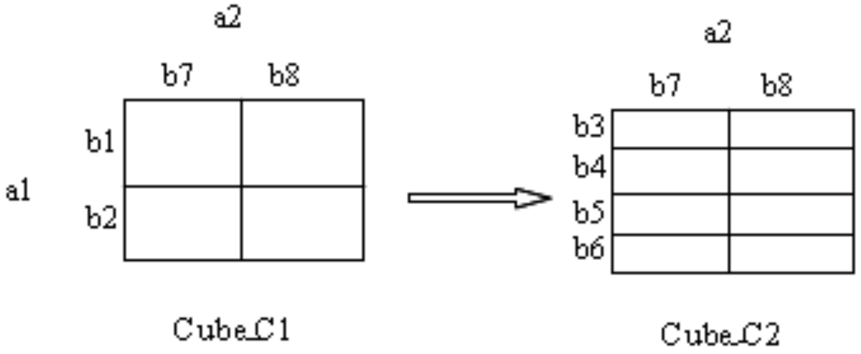


Figure 3: Drilling down in direction a1 - from Cube_C1 to Cube_C2.

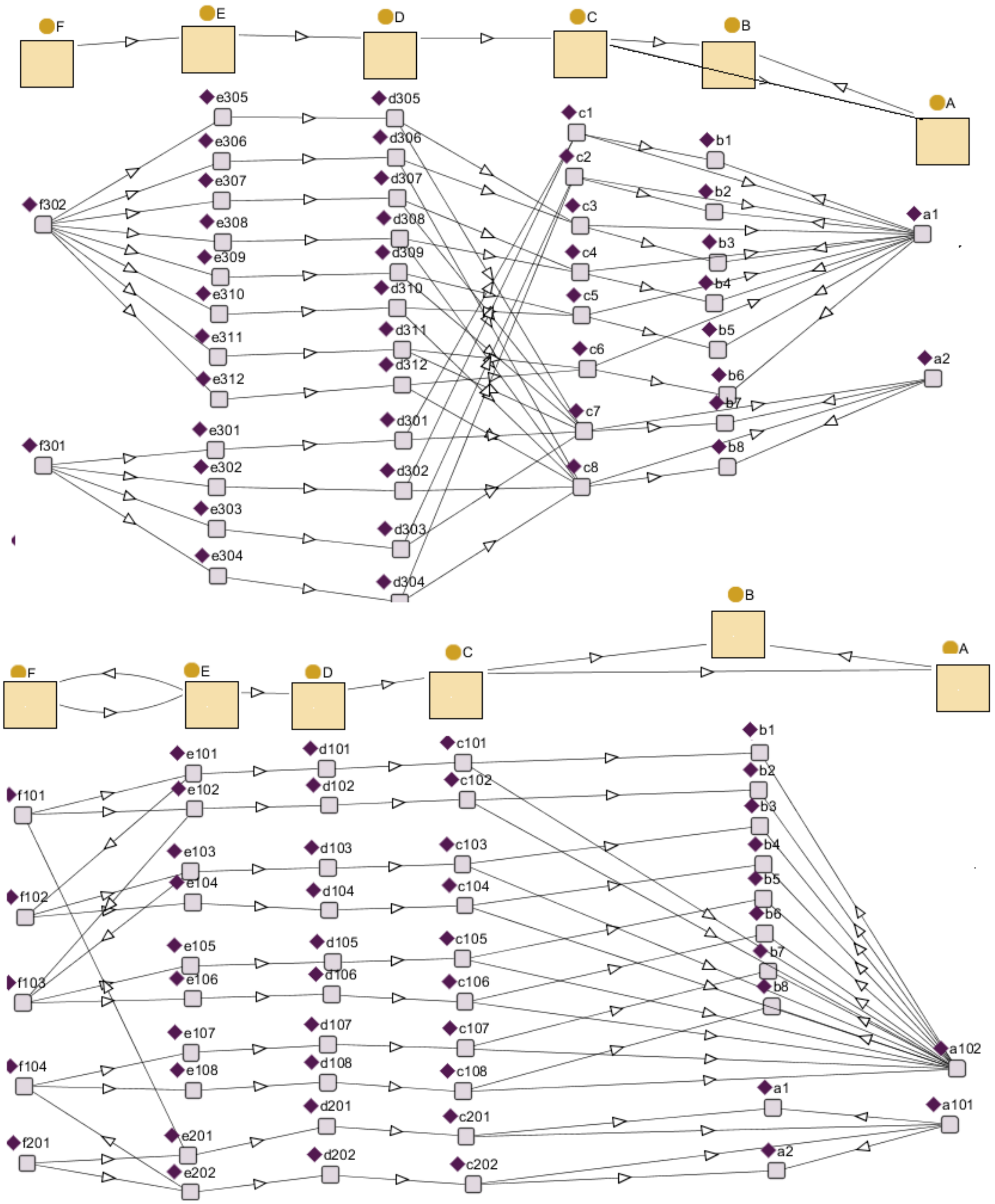


Figure 4: The OLAP_Ontology (top) and the Dimension_Ontology (bottom).

Dimension_Ontology. The instances b1, ..., b8 are instances of the class B in both. Order of f302, f301 in the OLAP_Ontology is saved by nesting instances f101, f102, f103, f104. The hierarchies of the dimensions a1, a2 are stored in the cells of f201. The f101 represents the first level of dimension a1 (levels b1, b2), the f102 and f103 the second level (levels b3, .., b6) , the f104 one level of a2 (b7, b8) (Figure **5**).

In P-ONT, operation of drilling down Cube_C1 in dimension a1 is performed as follows. First identify Cube_C1 as the instance f301 in OLAP_Ontology. For f301 choose instance a1. For a1 find instances of class B: b1 and b2. Go to the Dimensions_Ontology. In most external instance f201 of the class F, choose a1, and nested f101. For b1 and f102, read b3, b4. For b2 and f103, read b5, b6. Go to the OLAP_Ontology. Find

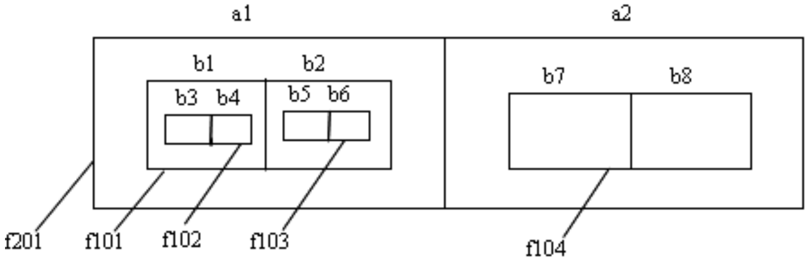


Figure 5: Scheme of nesting in the Dimension_Ontology.

instance of class F linked (indirectly) to b3, b4, b5, b6 instances of class B. Resulting f302 corresponds to Cube_ C2.

The OLAP operations slicing, collapse are realized programmatically basing SPARQL. Slice is a new cube with one fewer dimension, obtained by choosing a single level for one of its dimensions. The SPARQL query, parametrized by the dimension and the dimension level returns indexes of the new cube. The measures, the linked values from R, remain. Collapse is performed by the SPARQL queries giving slices. The measures got from the slices, linked to the same reduced by one index, are summed.

4. P-ONT BASED RELATIONAL DATABASE

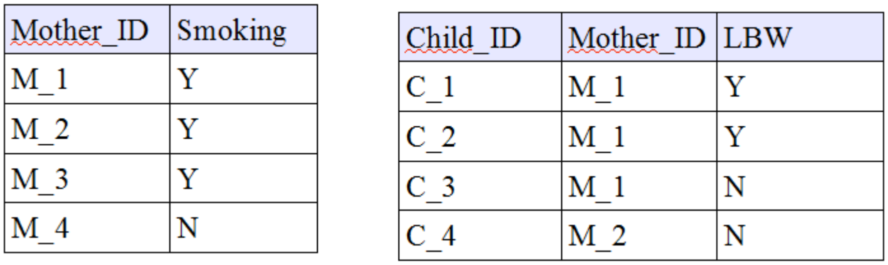
The problem of semantic mapping from databases to ontologies can be solved in different ways [20].

Our solution is creating relational database ontology in P-Ont. Consider two relational tables MOTHER and CHILD (Figure **6**). The table MOTHER has the primary key Mother ID and the column Smoking (Y/N). The table CHILD has primary key Child_ID, foreign key Mother_ID, one column LBW (Y/N).Records of the CHILD table are linked to records of the MOTHER table by foreign key Mother_ID.

The P-ONT ontology of relational database is expressed as a nested structure of cubes in the following order of nesting (Figure **7**). First level: the cube spanned on the columns of table MOTHER, here Smoking. Second level: the cube spanned on primary key Mother_ID. Third level: the cube spanned on the columns of table CHILD, here LBW. 4th level: the cube spanned on primary key Child_ID

Simple example is presented for the single relational table (Figure **8**).

The columns of relational table become instances of the Class AA and their levels become instances of



Mother_ID	Smoking
M_1	Y
M_2	Y
M_3	Y
M_4	N

Child_ID	Mother_ID	LBW
C_1	M_1	Y
C_2	M_1	Y
C_3	M_1	N
C_4	M_2	N

Figure 6: Relational tables MOTHER and CHILD.

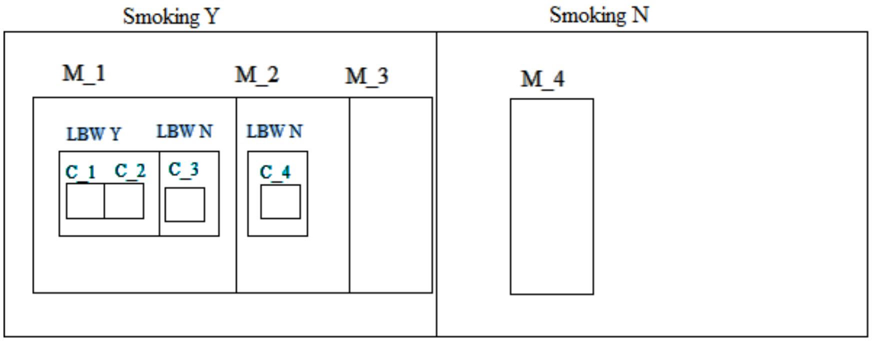
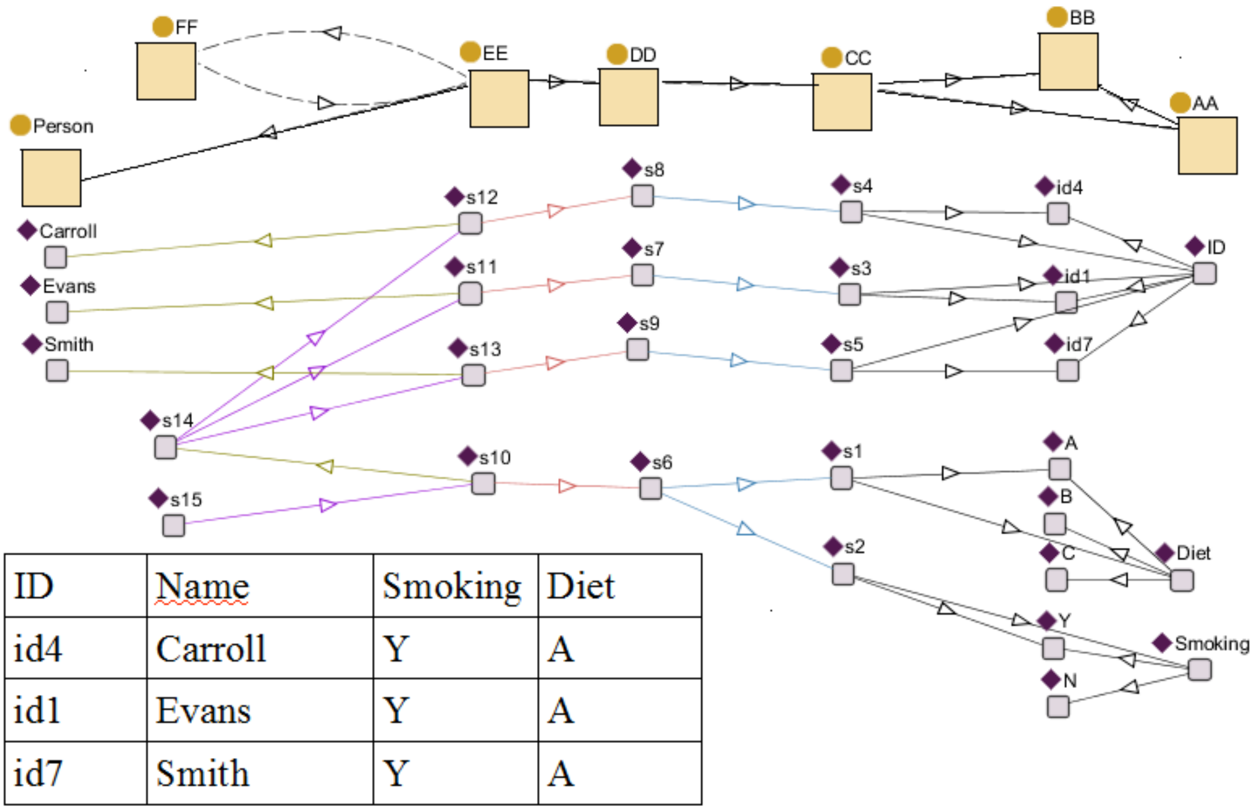


Figure 7: Scheme of Nesting Cubes of relational tables MOTHER and CHILD.



ID	Name	Smoking	Diet
id4	Carroll	Y	A
id1	Evans	Y	A
id7	Smith	Y	A

Figure 8: The ontology of the single relational table.

class BB. The relational table is written as two instances of class FF: s14 (dimension ID), and s15 (dimensions Diet, Smoking). Instance s14 is nested in s15 with index DietA; SmokingY. Data from relational tables are retrieved by the SQL query.

Data from P-ONT ontology of relational tables are retrieved by SPARQL query while maintaining the hierarchy of nesting.

5. P-ONT SEMANTIC WEB OF A STATISTICAL REGRESSION MODEL

5.1. Probability Concept in Comparison for Ontology

The immersion of probability into ontology is currently under consideration [21, 22]. Debates which are carried on probabilistic ontologies are conducted for the discrete case. This also relates to Bayesian networks identified with ontologies [23]. There is no fruitful studies on ontology concepts towards continuous variables.

In our understanding ontology describes the internal "architecture" of the phenomenon (relations between components of the phenomenon). Probability refers to the components itself not to the ontology. According to Kolmogorov, a probability space consists of three parts: given any set Ω (also called sample space); an algebra M on it; a measure P defined on M satisfying P(Ω)=1 (called a probability measure). It is a philosophical question to what extent the model reflects reality. Our pragmatic approach assumes that the "essential" knowledge of reality is not available. The model should present no more than you need (Occham razor).

We restrict ourselves to statistical linear regression models. This time Ω as a population is a mathematical image of a sets of statistical units existing in reality. Featured continuous attributes of the units (responses) has a probability distribution where the mean value vector is determined by the value of other attributes (predictors), variability of the responses is recorded by the matrix Var/Cov. For a single response matrix Var/Cov has one element. Predictor levels creates subsets so-called strata. The strata generates a -algebra M. Relations between strata are described by ontology.

Figure **9** shows strata designated by nominal predictors A, B, by jointly A and B and also by fixed values of the continuous predictor X. Arrows symbolize attached to the strata adequate parameters and fixed numbers. Thus, the model of the ANOVA can be

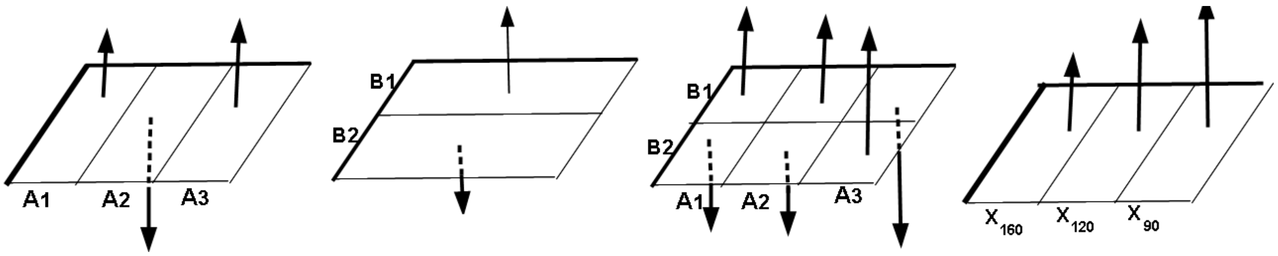


Figure 9: Strata for single nominal predictors, two-way ANOVA and strata designated by fixed predictor values of the linear regression.

identified with OLAP Cube where the stratum corresponds to the dimensions of the OLAP Cube, the parameters correspond to the measure.

In statistical analysis there are distinguished two aspects. The first aspect refers to the probabilistic models on the population (written by generalized linear models), the second refers to estimation of model parameters based on a sample, followed by hypothesis testing. In paper we consider estimation of model parameters and corresponding to that statistical model semantic web. Continuous predictors in such a situation have fixed values which denote the strata of the population. So, there is no need to distinguish between strata generated by nominal and continuous predictors. This means that the shape of the semantic web is independent of the measuring scale of the predictors.

5.2. P-ONT Semantic Web of the General and Generalized Linear Model

Let's start with a two-way ANOVA where: *Y* response *Neonate Birth Mass* is a continuous variable in population Ω. Nominal predictors of mothers are *A-Level of Blood Pressure* (*A_1-low; A_2-norm, A_3-hypertension); B-Smoking* (*B_1-Yes; B_2-No*). Parameters α_1, α_2, α_3 correspond to predictor levels *A_1, A_2, A_3*, parameters β_1, β_2 correspond to predictor levels *B_1, B_2* respectively. Moreover intercept is included (pooled mean value) in the model. The model is written as: $\mu_{ij} = \mu + \alpha_i + \beta_j$.

The model restrictions are: $\alpha_1 + \alpha_2 + \alpha_3 = 0$ and $\beta_1 + \beta_2 = 0$.

In the probabilistic interpretation μ_{ij} are mean values of the normal distribution *Y* given predictor levels α_i, β_j. An individual observation of neonate weight has value $Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$, where for any *ij* $e_{ij} \in N(0, \sigma^2)$.

Two-way ANOVA semantic web takes the form of two dimensional cube spanned on the directions A and B, where relevant cells contain model parameters α_i, β_j.

For ANOVA with interaction (first order), we have: $\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$. For the interaction there are the restrictions: $(\alpha\beta)_{11} + (\alpha\beta)_{21} + (\alpha\beta)_{31} = 0$ and $(\alpha\beta)_{21} + (\alpha\beta)_{22} + (\alpha\beta)_{23} = 0$. The semantic web is two dimensional cube spanned on the directions A and B, which relevant cells contain factors α_i, β_j and additionally $(\alpha\beta)_{ij}$. Note that the strata in population Ω are not explicitly shown in the linear equations. Figure **9** shows assigning values of the main components and interactions to strata (interaction parameters $(\alpha\beta)_{ij}$ are assigned to $A_i \cap B_j$ strata).

Let's go to the linear regression. Continuous response *Y* concerns to *Neonate Birth Mass*. Predictor X applies to the mother *Blood Pressure* is this time continuous also. Observed in the sample values of the *X* form a set of real numbers ($X_1, X_2, \ldots, X_k$). We assume X_1=160, X_2=120, X_3=90. The values of X_i designate the strata of women with this blood pressure value. Fixed part of the model takes the form: $\tau_i = Intercept + \alpha X_i$. Note that in ANOVA for nominal *Blood Pressure* there were three strata *A_1, A_2, A_3* and the corresponding parameters α_1, α_2, α_3. Now for the continuous value of *Blood Pressure* is one parameter α. Three strata (indexed by the numbers 160, 120, 90) correspond to values *X_1, X_2, X_3*. For individual mother with *Blood Pressure X_i*, the model is $Y_i = Intercept + \alpha X_i + e_i$, where $e_i \in N(0, \sigma^2)$.

Model restrictions requires that *Y* is normally distributed with the same variance in all strata and different mean values across the Ω strata.

From described above ANOVA and linear regression semantic webs result design matrices allowing estimation of the model parameters.

Let's consider now General Linear Models (GLM) in which predictors are nominal or continuous. GLM is presented as the linear equations (or the design matrix) [24-26].

Now consider the dependence of *Neonate Birth Mass* from cigarette *Smoking* (nominal) and maternal

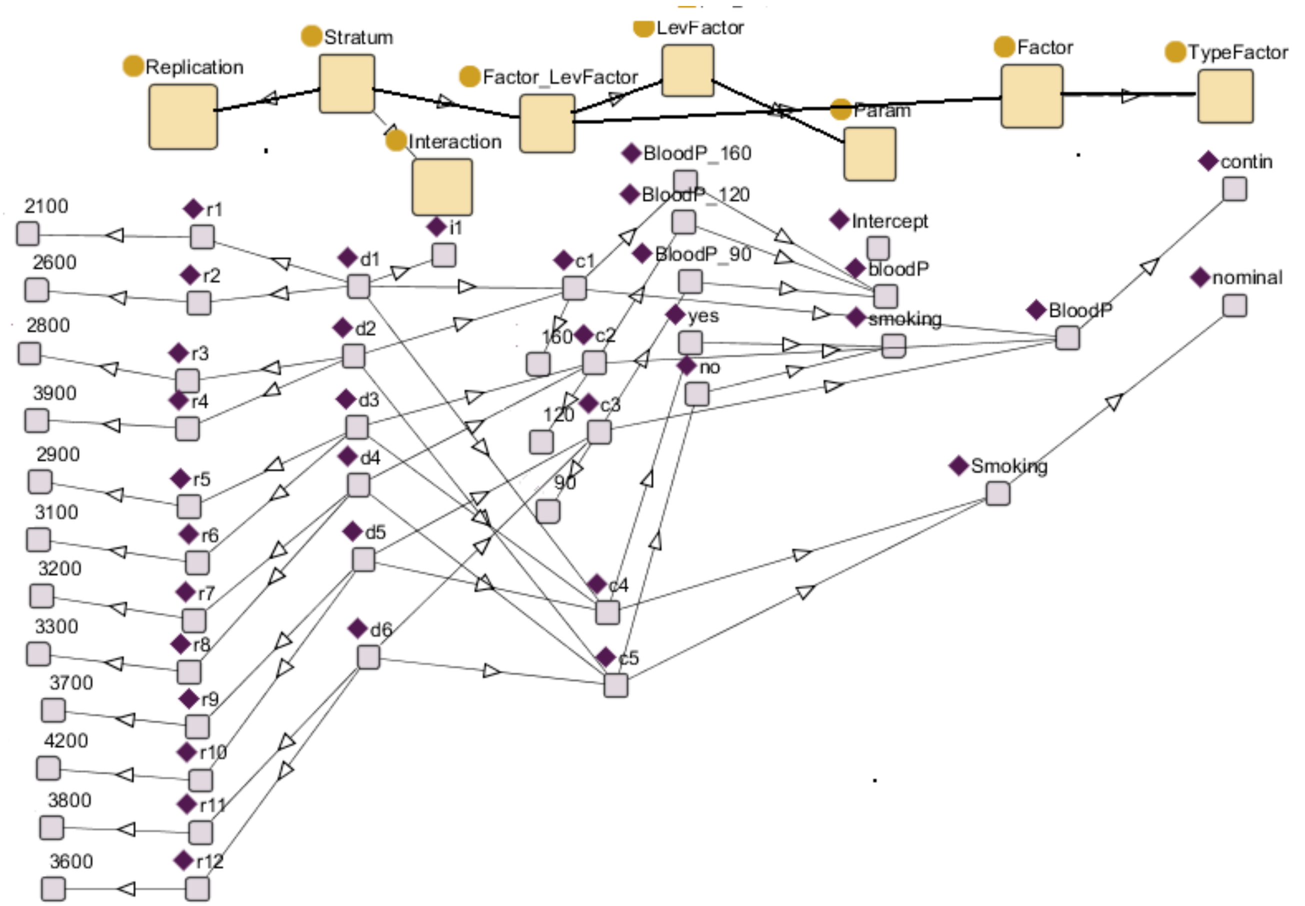


Figure 10: P-ONT Semantic web of the GLM- fixed part.

Blood Pressure (continuous). For clarity we assume that *Blood Pressure* takes in sample only three values: 160, 120, 90 (mm Hg). Is known that the same linear model can be written by different mathematical notation (different design matrix). In GLM commonly are used dummy variables for more than two nominal levels and multivariate linear regression (mutually exclusive) of the same continuous predictor.

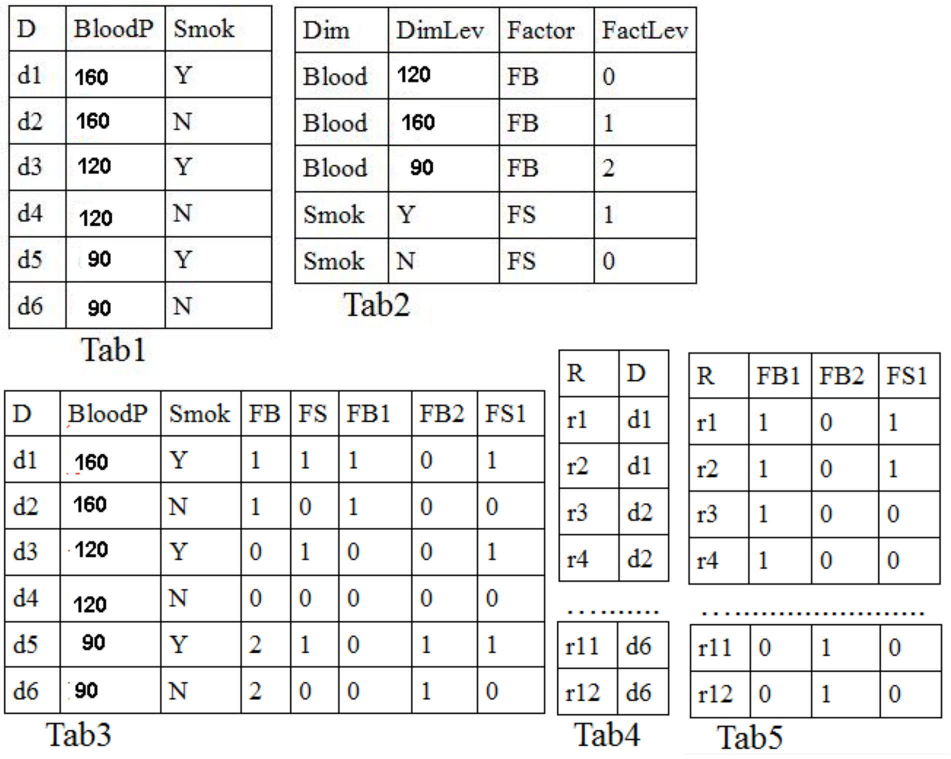
Let Y be response *Neonate Birth Mass* (Figure **10**).

Y_{ij} = *Intercept* + αX_i + β_j + $(\alpha\beta)_{ij}$ + e_{ij}, where X_1,=160, X_2=120, X_3=90 and where for any *ij* $e_{ij} \in N$ (0, σ^2).

Comment briefly presented above semantic web. In the Factor_LevFactor instances c1, c2, c3 correspond to levels BloodP_160, BloodP_120, BloodP_90, instances c3, c4 to levels SmokingYes, SmokingNo. Instances from Class Stratum correspond to strata BloodP*Smoking, instances from Param correspond to model parameters. Connection of a strata with model parameters provide links between instances of the Class Stratum and instances of the Class Param. Instances of the class Replication represent values of a Response and are associated by "data property" with real numbers. Instance i1 from Class Interaction correspond to interaction between BloodP and Smoking (for clarity interaction parameter in class Param is omitted).

Below is outlined an algorithm which makes the design matrix stipulated by the semantic web.

Algorithm for the GLM design matrix based on the P-ONT of the experiment (Figure **10**) constructs tables Tab1, Tab2, Tab3, Tab4, Tab5 (Figure **11**). SPARQL query returns instances of A class: BloodP, Smok. The Tab1 with columns D, BloodP, Smok is constructed. SPARQL query returns tuples of instances of the classes A(dimensions),B(dimension levels),D(indexes). In the D column of Tab1 are entered instances of the D class. In the columns BloodP, Smok are entered instances of the class B for BloodP and Smok. In Tab2 the BloodP and Smok (column Dim) with levels (column DimLev) are coded to the numerical factors FB, FS (column Factor) with levels (column FactLev). To Tab1 are appended columns FB, FS defined in Tab2 and the columns FB1, FB2, FS1 (Tab3). The



D	BloodP	Smok
d1	160	Y
d2	160	N
d3	120	Y
d4	120	N
d5	90	Y
d6	90	N

Tab1

Dim	DimLev	Factor	FactLev
Blood	120	FB	0
Blood	160	FB	1
Blood	90	FB	2
Smok	Y	FS	1
Smok	N	FS	0

Tab2

D	BloodP	Smok	FB	FS	FB1	FB2	FS1
d1	160	Y	1	1	1	0	1
d2	160	N	1	0	1	0	0
d3	120	Y	0	1	0	0	1
d4	120	N	0	0	0	0	0
d5	90	Y	2	1	0	1	1
d6	90	N	2	0	0	1	0

Tab3

R	D
r1	d1
r2	d1
r3	d2
r4	d2
....	
r11	d6
r12	d6

Tab4

R	FB1	FB2	FS1
r1	1	0	1
r2	1	0	1
r3	1	0	0
r4	1	0	0
....			
r11	0	1	0
r12	0	1	0

Tab5

Figure 11: Scheme of a design matrix creation.

FB1, FB2, FS1 which are 0/1 factors defined as follows. FB1=1 for FB=1; FB2=1 for FB=2; FS1=1 for FS=1; otherwise FB1, FB2, FS1 = 0. In statistical analysis FB1, FB2, FS1 are dummy variables, the value of zero is reference level.

The SPARQL query returns tuples of instances of R and D classes (Tab4). Combining Tab4 with Tab3 according to column D we received the design matrix (Tab5).

In the Generalized Linear Models (GGLM), in particular in the logistic regression, response takes in sample value 1 or 0, (Bernoulli distribution with parameter P). In our example, it is dependence of *Low Neonate Birth Mass (1-yes, 0-no)* from cigarette *Smoking* and maternal *Blood Pressure.* Model (without interaction) is $logit\ (ij) = Intercept + \alpha X_i + \beta_j + (\alpha\beta)_{ij} + e_{ij}$ Now the response is the *logit* (*ij*) (it contains "hidden" variability). Ontology for logistic regression is similar to the GLM model. In essence it is transformation of a probability response=1 into the odds $P_{ij}/(1-P_{ij})$ and subsequently transformation into the logit.

5.3. P-ONT Semantic Web of a Hierarchical General and Generalized Linear Model

One mother can give birth to twins or more babies. Thus a mother designates the cluster of neonates. For such a situation, it is appropriate hierarchical GLM and GGLM [27]. Consider the example of a two-level hierarchy, when baby is on the first level, mother is at second level. At first level *Neonate Birth Mass* is response Y, *Gender=(Male, Female)* with corresponding parameters γ_1, γ_2 is predictor of neonates.

Across clusters designated by the mother neonates mass mean value may vary. Thus, previously fixed parameters of the model changed to random variables. Conversion of fixed model parameters into the random variables we call randomization.

Smoking and *Blood Pressure* are the predictors of second level (mother) (observations of the *Blood Pressure* in sample are 160, 120, 90). Configuration of the strata of the second level hierarchy, is determined by the mother predictors. Second level ontology is analogous to the GLM ontology. The link between first and second levels we can imagine as multidimensional linear regression between the levels.

In mathematical notation, you can write it down as follows: First level is: $Y_i = Intercept + \gamma_1 + e_i$ (γ_2 omitted because $\gamma_1 + \gamma_2 = 0$) where $e_i \in N\ (0, \sigma^2)$. Connected together first and second level looks like this (shortened notation): $Intercept = Intercept_0 + \alpha_0 X_i + \beta_{0j} + \delta_0$; $\gamma_1 = Intercept_1 + \alpha_1 X_i + \beta_{1j} + \delta_1$

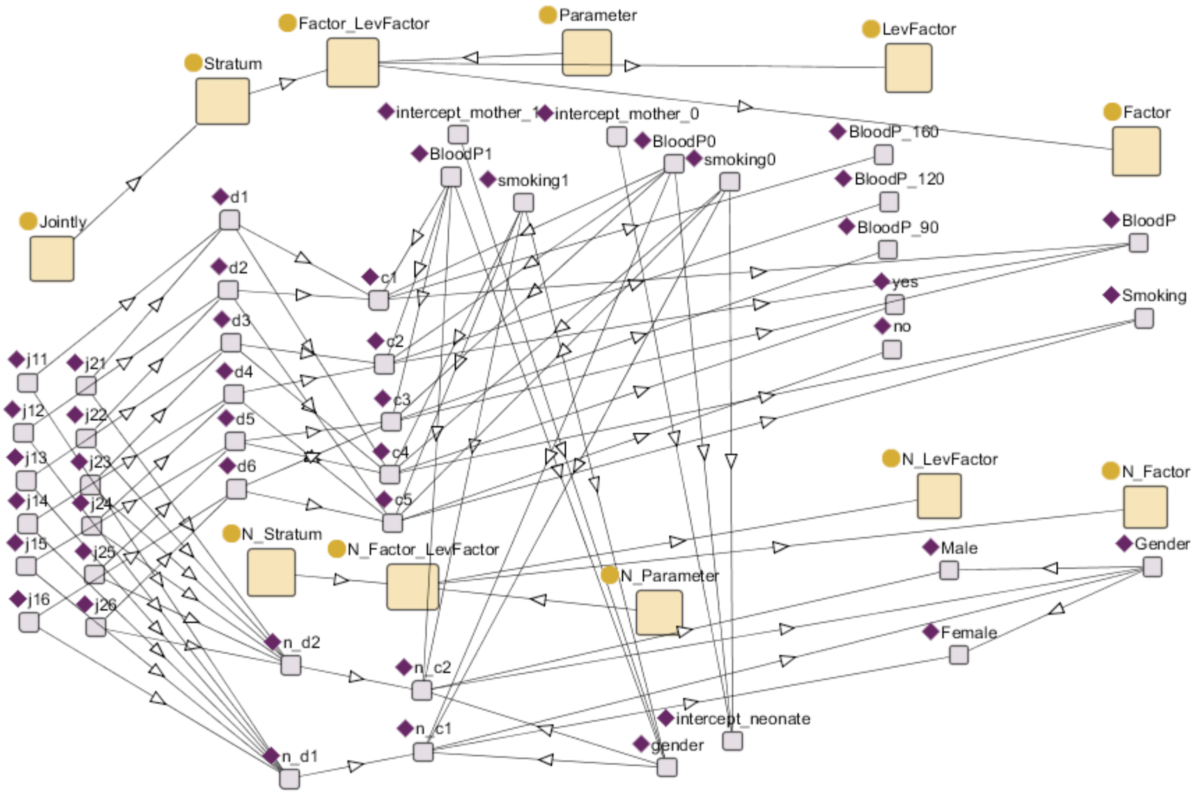


Figure 12: Semantic web of fixed part ot the hierarchical GLM.

Pair of numbers δ_1, δ_2 (in accordance with the assumptions) is derived from the two dimensional distribution N ((0, 0), Σ).

Semantic web of fixed part of the hierarchical GLM model is shown in Figure **12**. At the bottom of the figure is placed ontology for first level (neonate), at the top ontology for second level (mother). Ontology of the second level is analogous to Figure **10** with the difference that class Parameter contains duplicated parameters for intercept neonate and Gender form Class N_Parameter (parameters of the neonate). Blood pressure values present in the sample (160, 120, 90) are placed in a semantic web. Class Jointly contains instances corresponding to the strata of the mothers nested into strata of the neonates.

Figure **13** shows relation between model parameters on two levels and the connections to the random part of the model (Var/Cov matrix). Additionally neonates predictor not randomized (*Apgar*) is included in web.

Figure **14** shows clusters of neonates by mothers, needed to estimate the random effects.

Var/Cov matrix (*Σ*) is a generalization of the concept of the variance in the multidimensional case. Such a matrix for the random vector ($X_1, ..., X_n$) is shown in Figure **15-right**, where σ^2_i is the variance of the variable X_i, and σ_{ij} is the covariance between variables (X_i,X_j). In the P-ONT ontology Var/Cov matrix between the instances a1, ..., an of the class A is the instance of cube. The indexes of the cube are in the class D. There are introduced the auxiliary instances of A class axis1, axis2 and instances of the class B b1, ..., bn . The b1, ..., bn are equivalent to a1, ..., an. The axis1, axis2 (dimensions) are linked with b1, ..., bn (levels). The resulting cube, the set of instances of D class, is created for axis1 (b1, ..., bn), axis2 (b1, ..., bn). Figure **15** shows the ontology of the Var/Cov matrix of 2-dimensional normal distribution.

In the case of hierarchical GGLM on first level of the hierarchy is located ordinary GGLM.

6. STRATEGY OF AN AUTOMATED STATISTICAL INFERENCE

Due to the breadth and complexity of issues, the activities strategy is developed step by step. We test various options, including computer agent community (Jade platform) [28]. Work is in progress. Figure **16** gives an outline of our actions.

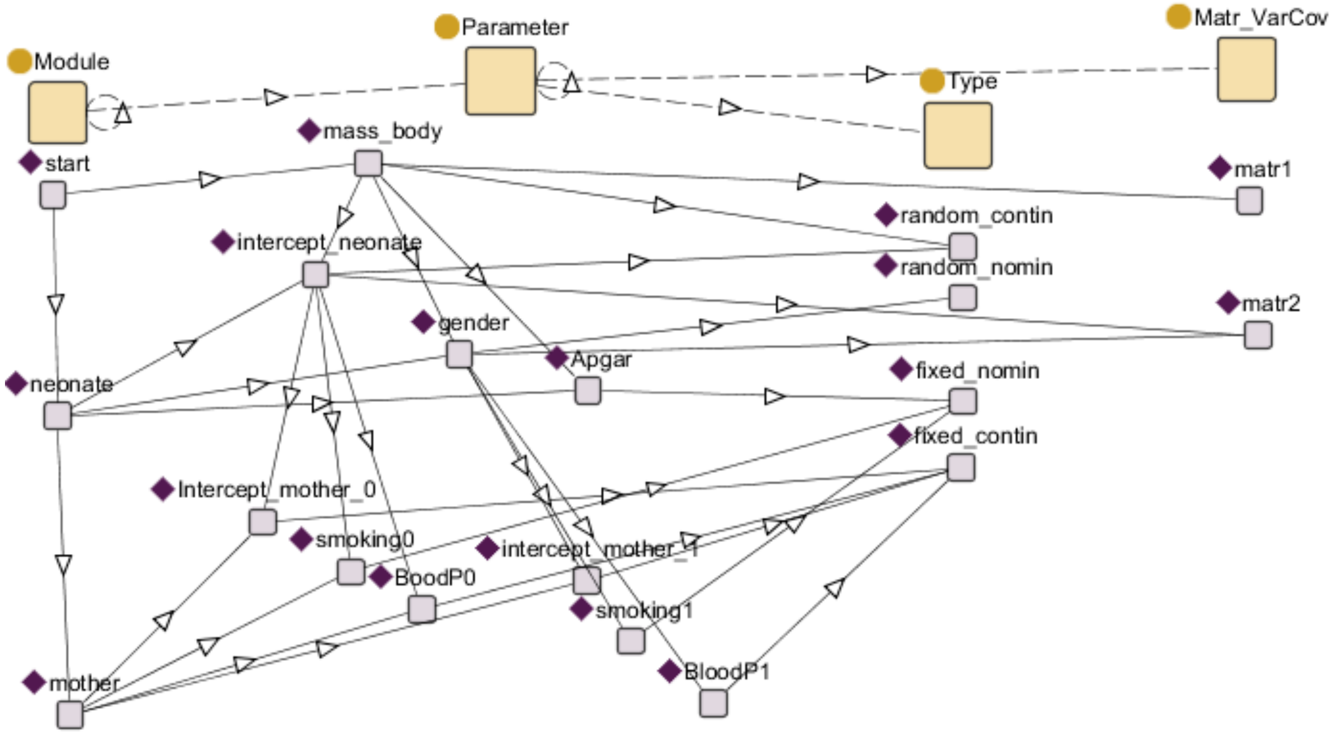


Figure 13: P-ONT of the hierarchical GLM-continuation.

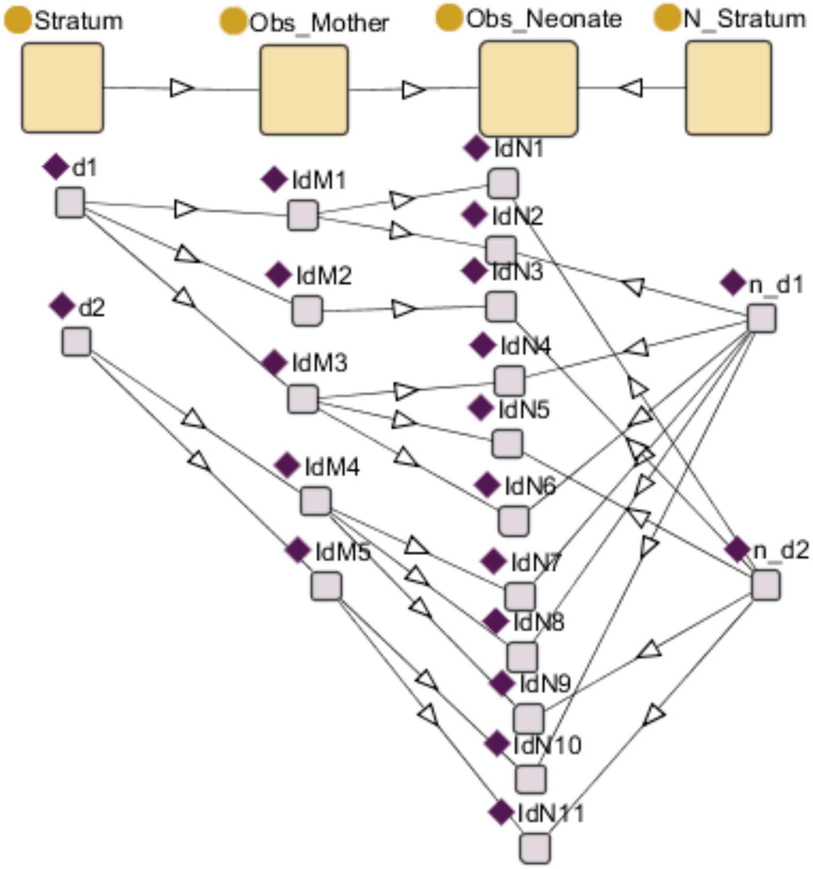
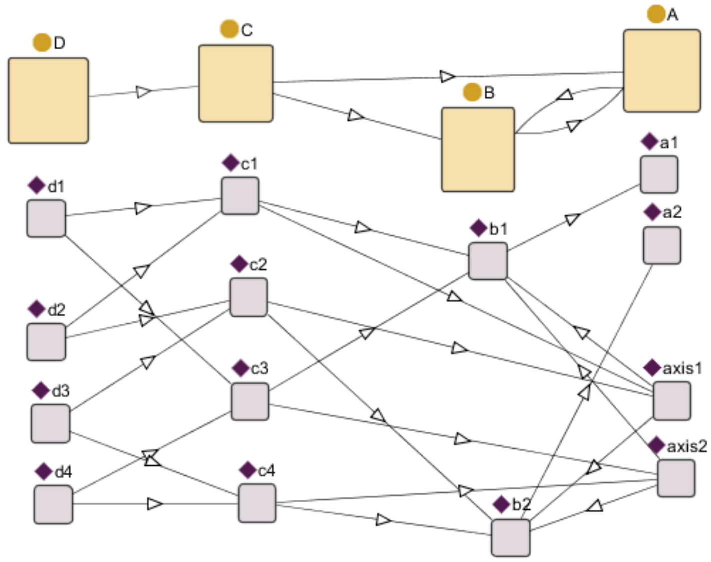


Figure 14: Module of the P-ONT hierarchical GLM with replications of the Mothers and Neonates.



$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \cdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix}$$

Figure 15: Semantic web of the Var/Cov Matrix for randomized parameters a1 and a2.

We start from experimental design. A statistical experiment is a methodical trial carried out with the goal of verifying a statistical hypothesis. Experiments provide insight into cause-and-efect by demonstrating what outcome occurs when a particular factor is manipulated.

In medicine are promising hierarchical mixed generalized linear models. In the web era observational studies are particularly important. Basic issue in observational study is the complex sampling (combined strata sampling with cluster sampling). The correct sampling provides the optimal size and balance of the sample. Sample size in the study is based on the expense of data collection, and need to have sufficient statistical power. Unbalanced sample causes the appearance of a confounding phenomenon (Figure **17**).

The table (cited in Wikipedia) shows the success rates and numbers of treatments for treatments involving both small and large kidney stones. Which treatment is considered better is determined by an inequality between two ratios (successes/total). The paradoxical conclusion is that open procedures are

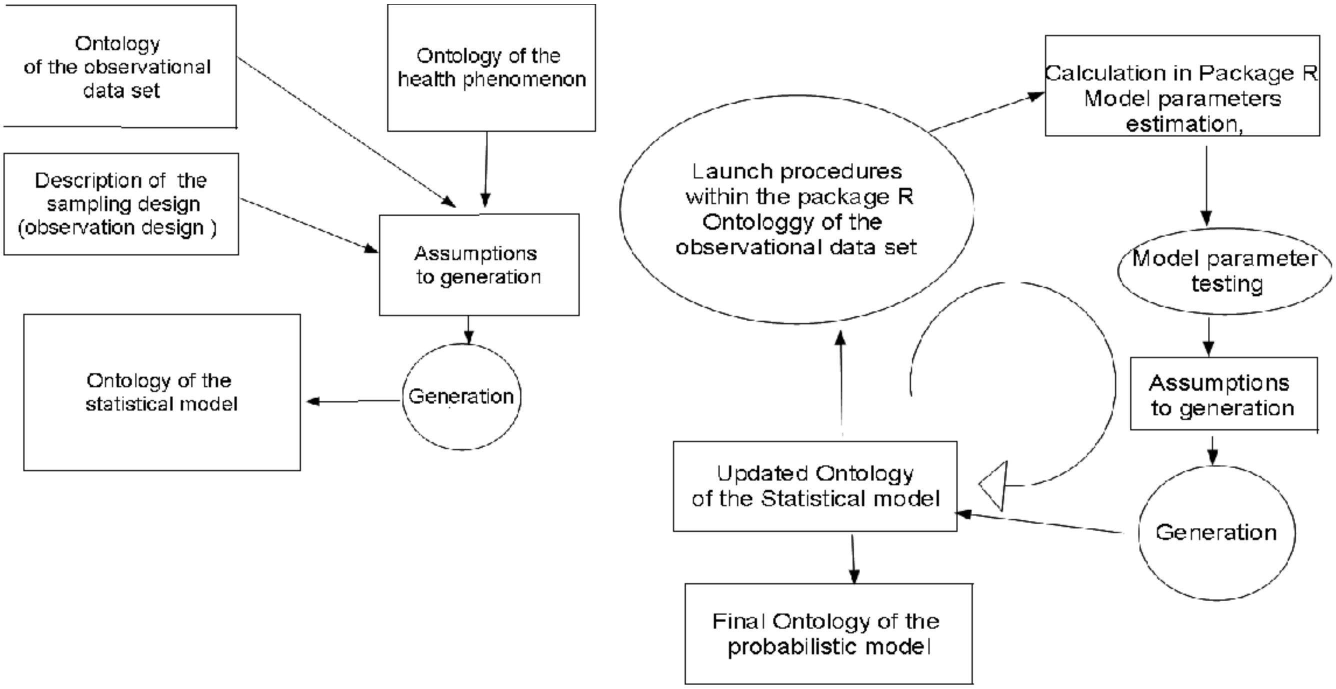
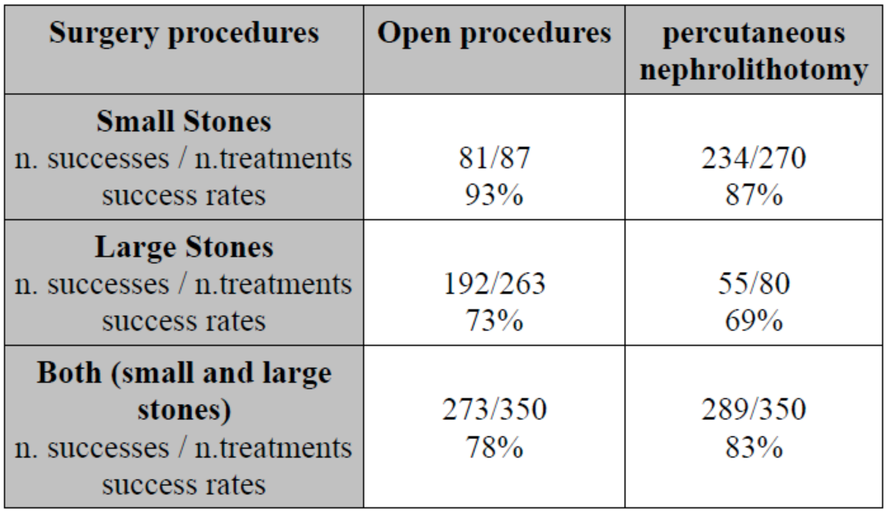

Ontology of the observational data set
Ontology of the health phenomenon
Description of the sampling design (observation design)
Assumptions to generation
Generation
Ontology of the statistical model
Launch procedures within the package R Ontologgy of the observational data set
Calculation in Package R Model parameters estimation,
Model parameter testing
Assumptions to generation
Generation
Updated Ontology of the Statistical model
Final Ontology of the probabilistic model

Figure 16: Course of the statistical inference process.



Surgery procedures	Open procedures	percutaneous nephrolithotomy
Small Stones n. successes / n.treatments success rates	81/87 93%	234/270 87%
Large Stones n. successes / n.treatments success rates	192/263 73%	55/80 69%
Both (small and large stones) n. successes / n.treatments success rates	273/350 78%	289/350 83%

Figure 17: Simpson Paradox example.

effective when used on small stones, and also when used on large stones. Percutaneous nephrolithotomy is more effective when considering both sizes at the same time. The confounding variable of the stone size was not previously known to be important until its effects were included.

A very important issue is the knowledge base that allow, inter alia, the selection of an appropriate statistical model.

According to common sense knowledge base is a database containing knowledge from specified domain. The problem is considered to be among the hardest of AI research because the breadth and complexity of knowledge is enormous. As a rule knowledge base is a type of ontology (written in descriptive language) of which the most general are called upper ontologies. In medical science there is used a knowledge base at a high level of maturity in broad range of subjects (UMLS, Gene Ontology) [4, 29], They are permanent and very important part of the research and diagnostics activity. Nevertheless, there is no good way to describe causal relationships and sequences of elementary facts that make up the whole process of pathogenesis. This opinion is based on review of medical bibliographic databases as well as mature studies of patho mechanisms of diseases [3, 30].

We have invented knowledge base as a system of causal quantitative relations described by real numbers. As the backbone of a health phenomenon knowledge base we adopted a hierarchical GGLM (in the form of P-ONT semantic web). It is an issue separate from the statistical inference, although closely connected with this. Work on it is still in progress.

7. COMPUTER IMPLEMENTATION

For illustration we will use ontology shown in Figure **2**. (in section 2.2). The GenMC generator creates the ontology of the multidimensional cube spanned on given dimensions. The file OWLEmpty.owl, the part of the GenMC, contains the Ontology of Cube with not specified dimensions, but with classes and properties. The generator gets from OWLEmpty ontology definitions of classes and properties and creates new ontology based on this. Parameters ,dimensions and dimension levels, for the generation of instances are taken from the Excel file. In neighbouring columns are saved dimensions along with their levels: Smoking-Y, Smoking-N, BloodP-L, BloodP-N, BloodP-H. Basing on them are created the instances of the OWL classes and linked by the OWL properties. From values of dimensions are created the instances of OWL class Dimension :"Smoking", "BloodP". From values of dimension levels are created the instances of OWL class DimLevel : "Y", "N", "L", "N","H". The DimLevel instances and the Dimension instances are linked by assigning levels to the dimensions in the input file. They formed instances of auxiliary class C: "Smoking.Y", "Smoking.N", "BloodP.L", "BloodP.N","BloodP.H". Linking the two instances of the class C, one for dimension creates instances of class D, indexes, "BloodP.L,Smoking.Y", "BloodP.L,Smoking.N", "BloodP.N,Smoking.Y", "BloodP.N,Smoking.Y ", " BloodP.H,Smoking.Y ", "BloodP.H,Smoking.Y ". The generator saves the just created ontology of two-dimensional cube of dimensions BloodP,Smoking in file BloodPSmoking.owl.

GenMC is java program. Input data saved in Excel file are read using Java Excel API. The OWL Ontology is created using Java OWL API. The OWL API provides java classes and methods to load and save OWL files, to query and manipulate OWL data models. Data model, instance JenaOWLModel java class, is defined by instances OWLNamedClass, OWLIndividual, OWLObjectProperty Java classes. Instances of OWLIndividual java class are created (method .createOWLIndividual) along with links between them (method.setPropertyValue). For generating ontologies of cubes of any number of dimensions with any number of levels dedicated computer programs are used. The GenMC generates such programs, which next generate the desired ontologies. GenMC is executed in steps, in which java code for consecutive step as well as fragment of final program is written. This tricky programming overcomes difficulty such as inability in Java of declaring new variable named by content of the string variable. String value of the variable is added to code of consecutive step, in which becomes the name of new declared variable.

This way of programming enables differentiating individual programs. Individual program contains "for" loops nested accordingly to number of dimensions, the names of dimensions are sewn into the program. Naming rules reflect the structure. The instance "BloodP.L,Smoking.Y"is linked with "BloodP.L"and "Smoking.Y", instance "BloodP.L" is linked with "BloodP" and "L", instance "Smoking.Y" is linked with "Smoking" and "Y". The naming rules, used in the generation of ontology process,however, are not necessary for reasoning , SPARQL selects by the structure. The java codes are processed as string variables and written to text files using methods of BufferedWriter class. The generated fragments are joined by methods of BufferedReader and BufferedWriter Java classes. The dimensionality of the cube does not enlarge computational complexity, nor the number of steps of the generation, but only the depth of nesting "for" loops. Building P-ONT ontologies we used the Protege-OWL editor designed for OWL format (W3C). Protege is a free, open-source platform, based on Java, with a suite of tools.

We widely used Jambalaya, a Protege plug-in to visualize Protege-OWL ontologies. Some Java applications in platform Eclipse were performed.

For reasoning, querying P-ONT ontologies, performing operations on OLAP Cubes we adopted an RDF query language SPARQL (W3C Recommendation).

For statistical calculations we use the R package. As an interface between R and Java language we employ r-java package [31].

8. CONCLUSIONS AND FURTHER WORK

Both epidemiological and statistical approach lead to a common goal. This goal is creating refined conceptual model of causation which will develop the already existing health phenomenon knowledge and discover new patterns revealing in random events and casuistic observations. It serves as a road sign, although we are aware that during the progress of research, it will recede, accordance with the rules of scientific research. Therefore, we are aiming at what can be achieved.

Currently we are working on demo application (written in Java) of automated statistical inference according to the scheme (Figure **16**). For that purpose we develop and test the generator of the hierarchical GGLM ontology. Its functioning is based on the principles which apply to the previously discussed OLAP Cube ontology generator. The data for generation obtained by SPARQL are placed in .xls file. They come from the health phenomenon knowledge base, semantic web of the observational data set (sample) and the description of the sampling design (necessary for determine the interaction parameters in the model). Moreover we prepare the generator, which creates matrix design tailored to the R Package syntax and the generator that returns a script triggering the complex sampling. We also are preparing applications that adjusts the existing SPARQL to P-ONT. In the P-ONT there are two levels of variables: classes as the first level variables, instances as the second level. So we are preparing P-ONT as two order language, which, as we believe, will improve the efficiency ontology generation process and increase unsatisfactory efficiency of SPARQL.

Innovation of our approach lies in the fact that the computer "understands" concepts written in logical language, as it was in OWL/RDF ontologies, and at the same time "understands" sense of the mathematical formulas necessary for the description and analysis observations from the real world. In the minds of many experts is vague conviction that an ontology is an explicit specification of a conceptualization (T. Gruber) [32]. In light of our work ontology is a formally written backbone (internal architecture) of a set of parameters and real numbers determining specific mathematical formulas.

REFERENCES

[1] Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. LIPPINCOTT WILLIAMS and WILKINS 2008.

[2] Perace N, Merletti F. Complexity, simplicity and epidemiology. Int J Epidemiol 2006; 35: 515-19. http://dx.doi.org/10.1093/ije/dyi322

[3] htpp://www.Cochrane.org/.

[4] Unied Medical Language System (UMLS), of Bethesda, MD: National Library of Medicine 2005. Springer 2005; vol. 2780-2003: pp. 51-60.

[5] Knublauch H. Protege-owl api programmers guide. Tech Rep, http://protegewiki.stanford.edu/wiki/ProtegeOWL API Programmers Guide/.

[6] Chen Y-J. Development of a method for ontology-based empirical knowledge representation and reasoning. Decision Support Systems 2010; 50: 1-20. http://dx.doi.org/10.1016/j.dss.2010.02.010

[7] Kleinberg S, Hripcsak G. A review of causal inference for biomedical informatic. J Biomed Inform 2011; 44: 1102-12. http://dx.doi.org/10.1016/j.jbi.2011.07.001

[8] Borkowski W, Mielniczuk H. Fcp based ontology of a statistical model for automated inference. 4-th International SWAT4LSW-2011 Workshop, December 2011.

[9] Borkowski W, Mielniczuk H. The use of cartesian product of owl classes to build a statistical model for the ai statistical analysis. ICB Seminar Statistics and Clinical Practice, International Centre of Biocybernetics-Poland Warsaw, June 2011.

[10] Klyne G, Carroll JJ. Resource description framework (rdf). Concepts and Abstract SyntaxW3C Recommendation. Tech Rep February 2004.

[11] Schneider P, Hayes P, Horrocs P. Owl web ontology language. semantics and abstract syntax. W3C Recommendation. Tech Rep February 2004.

[12] Hommeaux PE, S. A. Sparql query language for rdf. w3c recommendation 15. Tech Rep http://www.w3.org/TR/rdf-sparql-query/, January 2008.

[13] Guest and Editorial. Ontological foundations for biomedical sciences. Artic Intellig Med 2007; 3: 179-82.

[14] Sowa JF. Knowdge Representation: Logical, Philosocal and Computational Foundations, Cole Publishing Co., Pacic Grove, CA, 2000.

[15] Schulz S, Hahn U. Towards the ontological foundations of symbolic biological theories. Artic Intellig Med 2007; 39: 237-50. http://dx.doi.org/10.1016/j.artmed.2006.12.001

[16] Mielniczuk H, Borkowski W. Cartesian product of owl classes-theoretical foundations and application in the construction of probabilistic models in the space rn. in ICB,. Seminar Statistics and Clinical Practice, International Centre of Biocybernetics -Poland, Warsaw, June 2011.

[17] Vassiliadis P. Modeling multidimensional databases, cubes and cube operations. In Proc. of the 10th SSDBM Conference 1998; pp. 53-62. http://dx.doi.org/10.1109/SSDM.1998.688111

[18] Benczur A, Molnar A. An extended partition model for generalized multidimensional data 2007.

[19] Maniatis OA, Maniatis A, Vassiliadis P, Skiadopoulos S, Vassiliou Y. Cpm: A cube presentation model for in DaWaK 2003, Prague, Czech Republic, September 2003; pp. 3-5.

[20] An Y, Mylopoulos J, Borgida A. Building semantic mappings from databases to ontologies. In Proceedings of the Twenty-First National Conference on Articial Intelligence (AAAI-06) (Boston M, Ed.) 2006.

[21] Lukasiewicz T, Straccia U. Description logic programs under probabilistic uncertainty and fuzzy vagueness. Int J Approximate Reasoning 2009; 50: 837-53. http://dx.doi.org/10.1016/j.ijar.2009.03.004

[22] Lukasiewicz T. Expressive probabilistic description logics. Artic Intellig 2008; 172: 852-83. http://dx.doi.org/10.1016/j.artint.2007.10.017

[23] Shen TH, Tarczy-Hornoch P, Detwiler LT, *et al*. Evaluation of probabilistic and logical inference for a snp annotation system. J Biomed Inform 2010; 43: 407-18. http://dx.doi.org/10.1016/j.jbi.2009.12.002

[24] Keinbaum D, Kupper L, Muller LA. Appled Regression Analysis And Other Multivarable Methods. Wadsworth, Belmont CA, 1980.

[25] Keinbaum D. Logistic Regression A self-Learrning Text. Springer Verlag New York 1994.

[26] Linhart H, Zucchini W. Model Selection. John Willey And Sons 1986. http://dx.doi.org/10.1007/BF02932566

[27] Pinheiro JC, Bates DM. Mixed Eects Models in S and S-Plus. Springer Science, Business Medi, 2004.

[28] JADE-Java Agent Development Framework http://jade.tilab.com/.

[29] Ashburner M, Ball CA, Blake JA. Gene ontology: tool for the unication of biology. Gene Ontol Consortium Nat Genet 2000; 25(1): 25-29.

[30] Libby P. Atherosclerosis: The new view, Scientic American, vol. May 2002, 2002. http://dx.doi.org/10.1038/scientificamerican0502-46

[31] The R Project for Statistical Computing- http//www.r-project.org/.

[32] Gruber TR. A translation approach to portable ontology specications, knowledge acquisition. Curr Issues Knowledge Modeling 1993; 5: 199-20. http://dx.doi.org/10.1006/knac.1993.1008

Received on 06-09-2012 Accepted on 03-10-2012 Published on 21-12-2012

http://dx.doi.org/10.6000/1929-6029.2012.01.02.06