Dmitry Mouromtsev, Cyril Pchenichniy, Dmitry Ignatov (Eds.)

MSEPS 2013 – Modeling States, Events, Processes and Scenarios

Workshop associated with the 20th International Conference on Conceptual Structures (ICCS 2013)
January 12, 2013, Mumbai, India

Volume Editors

Dmitry Mouromtsev
National Research University of Information Technologies, Mechanics and Optics, Saint Petersburg, Russia

Cyril Pchenichniy
Intellectual Systems Laboratory
National Research University of Information Technologies, Mechanics and Optics, Saint Petersburg, Russia

Dmitry I. Ignatov
School of Applied Mathematics and Information Science
National Research University Higher School of Economics, Moscow, Russia

Printed in National Research University Higher School of Economics.

The proceedings are also published online on the CEUR-Workshop web site in a series with ISSN 1613-0073.

Preface

Human reasoning uses to distinguish things that do change and things do not. The latter are commonly expressed in the reasoning as objects, which may represent classes or instances, and classes being further divided into concept types and relation types. These became the main issue of knowledge engineering and have been well tractable by computer. The former kind of things, meanwhile, inevitably evokes consideration not only of a "thing-that-changes" but also of "change-of-a-thing" and thus claims that the change itself be another entity that needs to be comprehended and handled. This special entity, being treated from different perspectives as event, (changeable) state, transformation, process, scenario and the like, remains a controversial philosophical, linguistic and scientific entity and has gained notably less systematic attention by knowledge engineers than non-changing things.

In particular, there is no clarity in how to express the change in knowledge engineering - as some specific concept or relation type, as a statement, or proposition, in which subject is related to predicate(s), or in another way. There seems to be an agreement among the scientists that time has to be related, explicitly or implicitly, to everything we regard as change - but the way it should be related, and whether this should be exactly the time or some generic property or condition, is also an issue of debate.

To bring together the researchers who study representation of change in knowledge engineering both in fundamental and applied aspects, a workshop on Modeling States, Events, Processes and Scenarios (MSEPS 2013) was run on 12 January, 2013, in the framework of the 20th International Conference on Conceptual Structures (ICCS 2013) in Mumbai, India. Seven submissions were selected for presentation that cover major approaches to representation of the change and address such diverse domains of knowledge as biology, geology, oceanography, physics, chemistry and also some multidisciplinary contexts.

Concept maps of biological and other transformations were presented by Meena Kharatmal and Nagarjuna Gadiradju. Their approach stems from conceptual graphs of Sowa and represents the vision of change as a particular type of concept or, likely, relation, defined by meaning rather than by formal properties.

The work of Prima Gustiene and Remigijus Gustas follows a congenial approach but develops a different notation for representation of the change based on specified actor dependencies in application to business issues concerning privacy-related data.

Nataly Zhukova, Oksana Smirnova and Dmitry Ignatov explore the structure of oceanographic data in concern of opportunity of their representation by event ontologies and conceptual graphs. Vladimir Anokhin and Biju Longhinos examine another Earth science, geotectonics, and demonstrate that its long-lasting methodological problems urge application of knowledge engineering methods, primarily engineering of knowledge about events and processes. They suggest a

draft of application strategy of knowledge engineering in geotectonics and claim for a joint interdisciplinary effort in this direction.

Doji Lokku and Anuradha Alladi introduce a concept of "purposefulness" for any human action and suggest a modeling approach based on it in the systems theory context. In this approach, intellectual means for reaching a purpose are regarded either as structure of a system, in which the purpose is achieved, or as a process that takes place in this system. These means are exposed to different concerns of knowledge, which may be either favorable or not to achieving the purpose. The resulting framework perhaps can be described in a conceptual-graph-related way but is also obviously interpretable as a statement-based pattern, more or less resembling the event bush (Pshenichny et al., 2009).

This binds all the aforementioned works with the last two contributions, which represent an approach based on understanding of the change as a succession of events (including at least one event), the latter being expressed as a statement with one subject and finite number of predicates. The method of event bush that materializes this approach, previously applied mostly in the geosciences, is demonstrated here in application to physical modeling by Cyril Pshenichny, Roberto Carniel and Paolo Diviacco and to chemical and experimental issues, by Cyril Pshenichny. The reported results and their discussion form an agenda for future meetings, discussions and publications. This agenda includes, though is not limited to,

- logical tools for processes modeling,
- visual notations for dynamic knowledge representation,
- graph languages and graph semantics,
- semantic science applications,
- event-driven reasoning,
- ontological modeling of events and time,
- process mining,
- modeling of events, states, processes and scenarios in particular domains and interdisciplinary contexts.

The workshop has marked the formation of a new sub-discipline in the knowledge engineering, and future effort will be directed to consolidate its conceptual base and transform the existing diversity of approaches to representation of the change into an arsenal of complementary tools sharpened for various spectral regions of tasks in different domains.

January 12, 2013
Mumbai, India

Dmitry Mouromtsev
Cyril Pshenichny
Dmitry Ignatov

Organization

Workshop Co-Chairs

Dmitry Mouromtsev	National Research University of Information Technologies, Mechanics and Optics, Saint Petersburg, Russia
Cyril Pchenichniy	National Research University of Information Technologies, Mechanics and Optics, Saint Petersburg, Russia

ICCS Workshop Chair

Dmitry I. Ignatov	National Research University Higher School of Economics, Moscow, Russia

Program Committee

Paolo Diviacco	Istituto di Oceanografia i Geofisica Sperimentale, Italy
Tatiana Gavrilova	Saint Petersburg State University, Russia
Nagarjuna G.	Homi Bhabha Centre for Science Education, TIFR, Mumbai, India
Xenia Naidenova	Research Centre of Saint Petersburg Military Academy, Russia
Heather D. Pfeiffer	Akamai Physics, Inc., USA
Michael Piasecki	The City College of New York, USA
Jonas Poelmans	Kathoelike Universiteit Leuven, Belgium
Yuri Zagorulko	Ershov Institute of Informatics Systems (IIS), Siberian Branch of the Russian Academy of Sciences, Russia

Organizing Institutions

Homi Bhabha Centre for Science Education, Tata Institute of Fundamental Research, Mumbai, India
National Research University of Information Technologies, Mechanics and Optics, Saint Petersburg, Russia
National Research University Higher School of Economics, Moscow, Russia

Table of Contents

PurposeNet: A Knowledgebase Organized Around Purpose

Rajeev Sangal, Soma Paul, P. Kiran Mayee

Language Technologies Research Centre
International Institute of Information Technology
Hyderabad, India

sangal@iiit.ac.in, soma@iiit.ac.in,
kiranmayee@research.iiit.ac.in

Abstract. We show how *purpose* can be used as a central guiding principle for organizing knowledge about artifacts. It allows the actions in which the artifact participates to be related naturally to other objects. Similarly, the structure or parts of the artifact can also be related to the actions.

A knowledgebase called *PurposeNet* has been built using these principles. A comparison with other knowledebases shows that it is a superior method in terms of coverage. It also makes it possible for automatic extraction of simple facts (or information) from text for populating a richly structured knowledge-base.

An experiment in domain-specific question-answering from a given passage shows that PurposeNet used alongwith scripts (or knowledge of stereotypical situations), can lead to substantially higher accuracy in question answering. In the domain of car racing, individually they produce correct answers to 50% and 37.5% questions respectively, but together they produce 89% correct answers.

Keywords: Ontology, Semantic Knowlegdebase, Information Extraction, OWL, Question-Answering

1 Introduction

There is a need to represent knowledge for a variety of applications, ranging from natural language processing to reasoning in sciences, education, business, social science and humanities. This requires Knowledge Representation (KR) schemes, as well as good ways of organizing knowledge.

KR schemes and inference methods have received a great deal of attention. This has resulted in several effective schemes which are strong as well as have efficient and powerful inference methods. Notable among them have been Sowa (1984), (2002), (2005) and Bharati et. al (1987), (1991), (1995).

Besides the KR schemes, there is also a need to work out the organization of knowledge. The question naturally arises as to what principles should be used to organize knowledge, namely, what knowledge should be put in, and how would parts of

that knowledge relate with other parts of knowledge? For example, if the domain of transport needs to be described, how should the different elements starting from car and trucks and going to repairs and roads, be organized?

The answer lies in recognizing that there are principles underlying the organization. Once these are understood, it becomes easier to relate different parts of knowledge with each other. Such knowledge can then be represented in a suitable KR scheme.

We have used purpose as an organizing principle in our work. This principle has been applied primarily to artifacts or manmade objects. It has been developed and used extensively in the Indian philosophical tradition. Objects are described in terms of four major types of attributes: rup, gun, svabhav, dharm.

Dharm is that property which is intrinsic (essential) to the objects in the category, and helps distinguish the category from other categories. Dharm is given by its purpose. For example, for a car, its dharm or purpose would be to transport (a small number of) people from one place to another on land.

Svabhaav refers to those attributes which the object shares with objects of the same class and which it does not share with other classes. For example, Car shares attributes with other machines, but does not share attributes with living beings.

Rup (literally meaning, form) refers to those attributes which can directly be perceived by our sensory organs. For example, rupa of car would be its shape, colour, weight, etc. Gun refers to properties that are not perceived directly but indirectly such as load carrying capacity, etc. dharm and gun are performative, where assvabhav and rup are non-performative (though they are essential for performance).

While building PurposeNet, a knowledgebase, we have used purpose as the primary principle of organizing knowledge. We note that the dictionary uses the same idea to give meanings of words. Let us take some examples from popular resources such as WordNet (Miller et. al, 1990), Wikipedia (Wikipedia, 2004) and Cambridge dictionary (http://dictionary.cambridge.org/ dictionary/american-english/).

WordNet defines the artifacts “fork”, “bomb” and “knife” in the following manner:

1. Fork - cutlery used for serving and eating food.
2. Bomb - an explosive device fused to explode under specific conditions.
3. Knife - edge tool used as a cutting instrument; has a pointed blade with a sharp edge and a handle.

In Wikipedia articles on artifacts, the first sentence generally describes the artifact as exemplified below:

1. Chair – A chair is a raised surface, commonly for use by one person.
   Wall – A wall (from Old English weall) is a vertical structure, usually solid, that defines and sometimes protects an area.
2. Football - A football is an inflated ball used to play one of the various sports known as football.

Cambridge dictionary has the following entries:

1. Telephone – A device for speaking to someone in another place by means of electrical signals
   Brush – Any of various utensils consisting of hairs or fibers arranged in rows or grouped together, attached to a handle, and used for smoothing the hair, cleaning things, painting, etc.
   Rack – A frame, often with bars or hooks, for holding or hanging things.

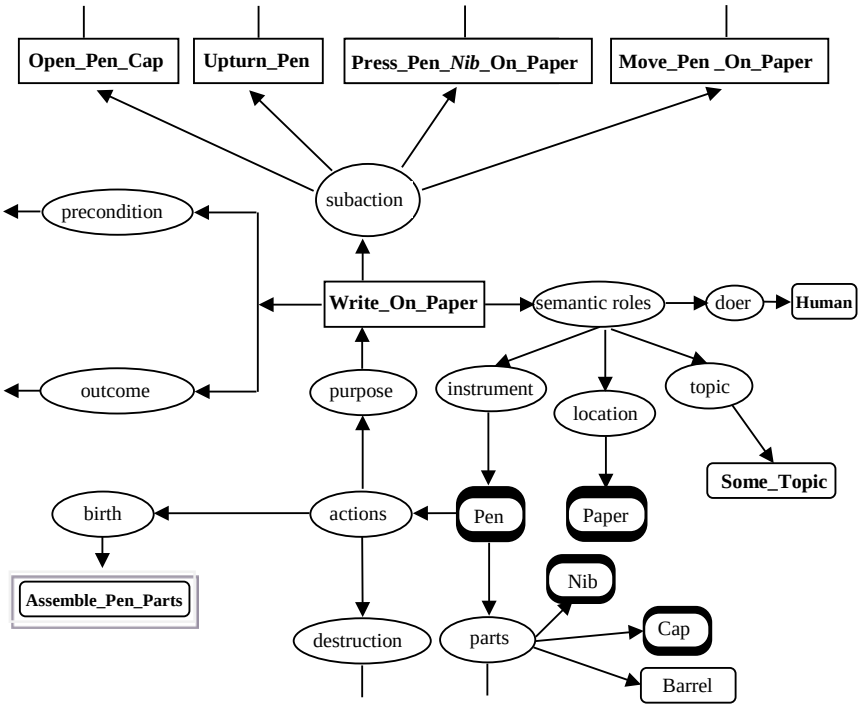


Fig.1. Illustration of importance of Purpose as a basis for knowledge representation.

All the nine entries cited above are defined in the form "an *X*<is a *Y*><with purpose *Z*>", where, *X* ={Fork, Bomb, Knife, Chair, Wall, Football, Telephone, Brush, Rack}, *Y* = {cutlery, explosive device, edge tool, raised surface, vertical structure, ball, device, utensil, frame}, *Z*={to sit on, that protects an area, to play, for throwing, for holding, …}. Thus, purpose is very significant information about artifacts. An artifact is made in order to serve one particular purpose. The various characteristics and activities associated with an artifact depend upon the purpose for which it is created.

As one would have noticed, the purpose of an object is given in terms of an action that the object helps accomplish. The object also has a structure, i.e., is made up of parts which are put together in well-defined way. The structure is related to the purpose of the object, namely, the structure helps accomplish the purpose.

In the case of a pen, for example, the purpose is to write on paper. The Pen has a thin and cylindrical shape for a comfortable gripping while used for writing. It has many sub-parts, such as Barrel, Nib, Feed and Cap, which together help carry out the action of putting marks on paper. The action can be broken into sub-actions which relate to the parts, where each part helps in carrying out some sub-action(s). Barrel holds Ink, Nib allows Ink to pass through and Cap prevents the Ink from drying. Therefore, when Ink-Pen is made, it is an assemblage of the aforementioned components and we know why the components are in the way stated. Each of them helps in fulfilling the central purpose of Pen, which is, writing.

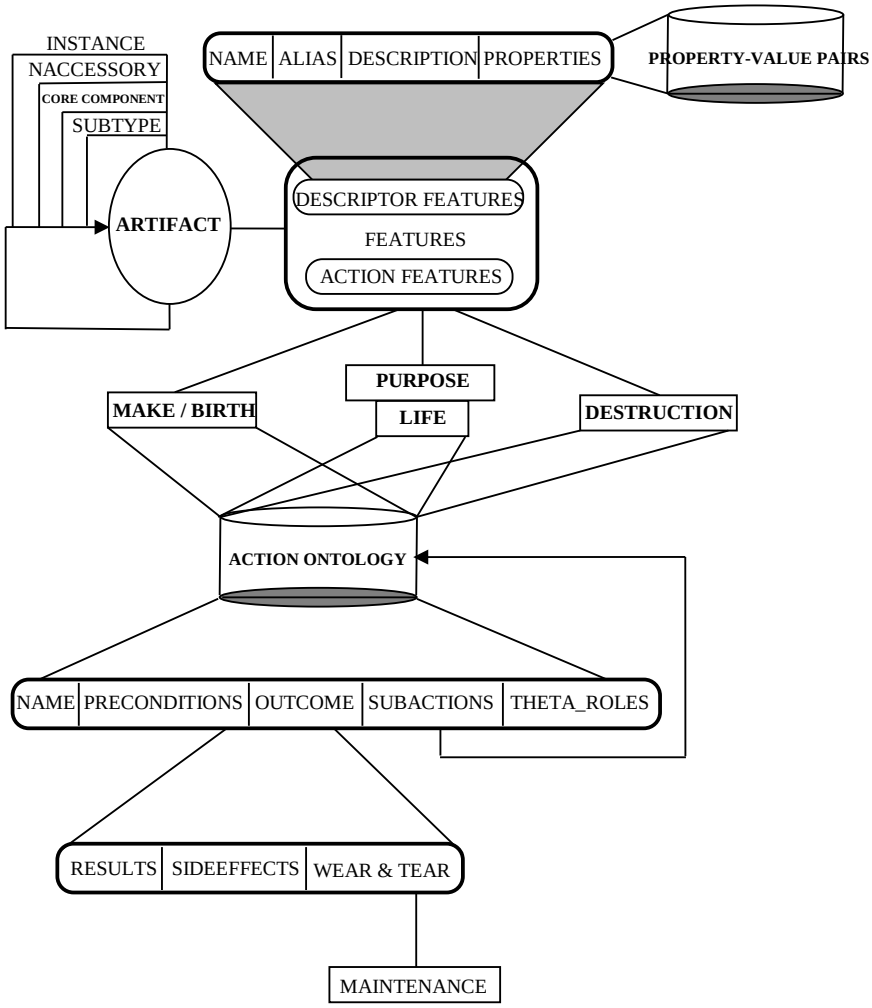


Fig. 2. Architecture of PurposeNet

If we look at the life cycle of an entity, we find that it has three major phases: creation, life and destruction. The purpose of an artifact is fulfilled at the second phase of life cycle, namely, when it has life. Therefore, at this phase, the artifact gets associated to other entities without which the purpose cannot be fulfilled. For example, a human being is an 'agent' who uses 'Pen' as an 'instrument' of writing. The writing is done on a smooth surface, for example, Paper. Ink is a requisite for writing. Thus, the artifact Pen is now related to the artifacts Ink and Paper as well as a 'human agent' without whom the action of 'writing' will not take place. There might be a change of state, for example, a Pen-Barrel can break; Ink gets over after a period of time. Finally, in the third phase of the life cycle of the artifact, it undergoes destruction. For example, 'Pen' undergoes destruction and gets converted to another entity, such as the reuse of metal parts for making of some other entity, such as 'Staple Pin'. It is therefore possible to engineer a knowledgebase of entities based on the characteristics activities and states of entities. Whereas object-oriented paradigm suggests that objects should be the central focus for engineering knowledgebase, our observations on entities suggest that entity-based knowledge cannot be complete unless it is focused on the purpose of entities and the actions that the artifacts are involved in.

We formally define PurposeNet in the following terms:

PurposeNet is a knowledgebase of artifacts with its properties, relationships and actions in which it participates with purpose as the underlying design principle.

2 Architecture of PurposeNet

PurposeNet has the artifact as its primary focus for organizing knowledge. Artifacts are fully described by its features and relationships with other artifacts. Two kinds of features have been postulated for the task: descriptor features and action features. The details of these features are given in section 2.1. Artifacts can also be described by the company it keeps, i.e. its relation with other artifacts as illustrated insection 2.2. The architecture of the PurposeNet is shown in the figure 2.

2.1 Features

The various distinct properties of an artifact are called its features. These features may be morphological such as the physical state of the artifact, its size, shape, magnitude and so on. The features may also be physiological like make, wear and tear, activities it performs, and so on. Based on whether the feature is morphological or physiological, we subcategorize features into the *descriptor features* and *action features.*

Descriptor Features.

The descriptor features of PurposeNet have three constituents that are found in WordNet as well, viz., Name, Alias and Description. SUMO has one attribute *Internal* that contains some properties which are similar to PurposeNet *descriptor features*. However SUMO properties are limited to olfactory, visual, texture and taste, with no

further refinement. The descriptor features of PurposeNet have been prepared after a study of texts of *Nyaya-Vaisheshikadarshana*(Prasastipada(1977), Singh(2001),Kulkarni(1994)) and others (Isvarkrsna(2007), Nagaraj(2003), Cowell(2001)).

Descriptor Feature	Definition	Value
Color	The property possessed by an object of producing different sensations on the eye as a result of the way it reflects or emits light	Red, Blue, Green, Cyan,Indigo, Orange, Pink, Black, White, Any
Constitution	The material with which an artifact is made of	Metal, rubber, wood, foam, plastic, glass
Shape	The external appearance of an artifact	Cubical, Oval, Triangular, Circular, Spherical, Aero, any
Size	The amount of space occupied by the artifact	Microscopic, very small, small, medium, large, any
State	The physical state in which the artifact usually exists	Solid, liquid, gas

Table 1. Descriptor Features and their description

From the complex set of properties, we have selected twenty five based on the ones most suitable for all artifact types. Also, we have added properties of significance such as Standard Capacity, Standard Weight, and, Physical State to enable a more comprehensive representation of information about artifacts. The possible values that can be taken by these properties (qualitative) have been extracted from various sources, including Wikipedia, Alani and Brewster(2006), Helmhotz (1970), Sunder Rao (2003), and Gayatri Devi (2007). A brief description of some properties in *descriptor features* is given in table 1. Comprehensive Descriptor feature list is given in Appendix.

The value of some descriptor properties with respect to the artifact Car is given in table 2:

SNo	Decriptor_Feature	Value
1	Name	Car
2	Alias	Automobile
3	Description	A type of motor vehicle used to transport people.
4	State	Solid
5	Shape	Aerodynamic

6	Color	Any
7	Constitution	Metal
8	Size	Moderate_Size

Table 2. Values of Descriptor properties for the artifact Car

Action Features

Since the very need for an artifact is to serve some purpose in the human environment, it is understood that every artifact is associated with some actions. The various activities associated with an artifact constitute its Action Features. This categorization has been developed based on the various stages in the Lifecycle of an artifact. The first stage of an artifact is its Make or Birth. It is then prepared for the first-time use, after which it reaches the purpose-serving stage, i.e., Life. Here it may be prepared again for reuse or may be in the general or repair-related maintenance stage. From here, the artifact again goes back to the purpose-serving stage. After one ormore iterations of the purpose-serving stage, the artifact becomes no longer usable, which is when it is in the Destruction stage, and is therefore a last stage activity. Its individual parts are recycled and it becomes the basis for the birth of the same or another category of artifact. The various action features are accordingly classified primarily as – make actions, purpose-serving actions, and, actions after destruction. The secondary actions are first-time preparation-before-use actions that makes an artifact usable and the trio of subsequent preparations before use actions, general maintenance and repair maintenance actions that allow for subsequent usage of an artifact. Table 3 shows these actions for Transport_using_Car artifact.

SNo	**Action Feature**		**Value**
1	Make/Birth		1. Integrate(Car_Interior_Parts) 2. Integrate(Chassis and Car_Body)
	1a.	First-time-Preparation before use	1. Fill(Car_Fuel) 2. Test(Car_Pedals) 3. Test-Drive(Car) ….
2	Life - Purpose		Transport things
	2a.	Subsequent preparation before use	1. Check(Fuel) 2. Test(Car_Pedals) 3. Check(Rear_View_Mirror) …...
	2b.	Repair Maintenance	1. Repair(Car_Engine) 2. Repair(Car_Ignition_system) 3. Repair(Car_Pedals) 4. Repair(Car_Door) ….
	2c.	General Maintenance	1. Wash(Car) 2. Oil(Car_Engine)

			3. Oil(Ignition_System) 4. Fill(Car_Tyre)
3	Destruction		1. Car_Engine - Recycled-to-metal 2. Car_Tyre - Recycled-to-fuel-and-oil 3. Car_Chassis - Recycled-to-another-Car 4. Car_Seat - Reused-in-another-Car

Table 3. Table showing all the Action-features of a Car

Every non-primitive action can be fully described using a quadruple consisting of its preconditions, outcomes, subactions and semantic roles. We call this Quadruple as the action frame. Every primitive action can be described using the same frame as above, minus the subactions. This description remains unchanged irrespective of the broad category into which the action belongs – i.e., whether it is birth or make action, or action related to life. The action frame places a formal structure on the Action features (Kiranmayee et al., 2011).The action frame for a sample action, namely 'transport thing', which is the purpose of the artifact Car is given in table 4.

Artifact: Car:: Purpose – Transport_Thing			
No	**Action Frame Element**		**Value(s)**
1	Precondition		1) Exists_Car_at_Source 2) Exists_Thing_Near_Car
2	Out-come	Result	1) Change_Position (Thing)
		Side Effect	1) Change_Position (Car) 2) Change_Position (Driver)
		Wear-and-tear	1) Wornout(Engine) 2) Wornout(Tyre)
3	Subactions		1) Load(Thing) 2) Drive(Car) 3) Unload(Thing)
4	Theta Roles		1) Theme – Thing 2) Source – Place 3) Destination – Place 4) All other Roles – Null

Table 4. The Action Frame for the action *transport*_thing_from_Source_To_Target

2.2 Relations

An artifact can also be described in terms of its association with other objects in the world. For example, objects that come to our mind when we think of the artifact Car might be the following: engine, wheel, steering, gear, seat, petrol, diesel, road,

petrol pump, car window, music system, rear view window, car body and so on. The relations of these artifacts with car exist at different planes in terms of purpose that the Car is used for. The primary purpose of Car as shown in table 3 is 'transport things from one place (*X*) to another place (*Y*)'. In order to fulfill the action of transporting, a Car needs to move from *X* to *Y* and we call the action 'drive'. For 'drive' action to take place, following parts of Car which claim to have a purpose of their own, is essential: Engine, Wheel, Steering, Gear. Such components are called Core Component. Rear view window is also part of Car but it is useful for some specific movement of car (i.e., when the car moves back). Such components are called purpose-serving-accessory in contrast to non-purpose-serving-accessory such as AC, music system which are parts of Cars but are not directly related to Car driving. Other kind of artifacts such as petrol, diesel, road are directly related to driving even though they are not part of Car. Such artifacts are related to Car with in terms of a relation called Naccessory. Apart from these relations, there exist the usual subtype relations between an artifact and its specific types. The following figure demonstrates various relations and example cases for the artifact Vehicle:

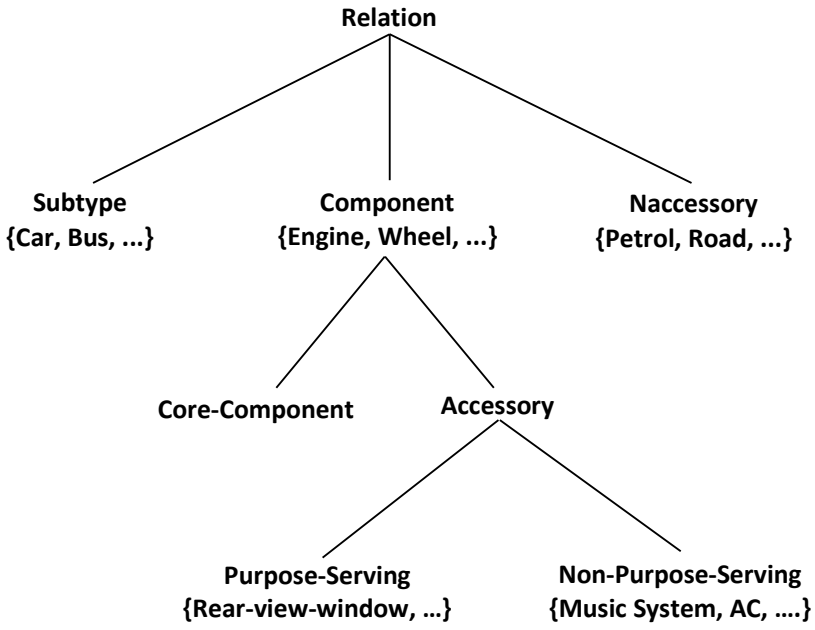


Fig.3. Relations describing an artifact

3 Implementation

The best possible design to represent our architecture of PurposeNet is object-oriented and top-down methodology. The PurposeNet knowledge base has been implemented using the concept of Ontology. Ontology is a formal explicit description of concepts in a domain of discourse, properties of each concept describing various features and attributes of the concept, and restrictions on slots. Ontology together with a set of individual instances of classes constitutes a knowledge base (Noy, 2001). Ontology helps us develop the Semantic Web, which is a vision for the future in which information is given explicit meaning on the web, making it easier for machines to automatically process and integrate information. We have chosen OWL to implement our knowledgebase.

3.1 Statistics of PurposeNet Implementation

The active ontology for purposenet in Transport domain has an Artifact count of 3678 (Car_Door, Car, Car_Hinge ...), general property count of 87 (Color, Shape, …, Birth, Processrel, ...), data property count of 8 (capacity, number, ...), Instances count of 264 (Audi_A4, BMW_6_Series, Chevrolet_Tahoe, Daewoo_Matix, ...), and Sub-Classes count of 8045 (Car_Rear_Seat, Car_Passenger_Seat, ...). The same is developed Semi-automatically by Domain Experts. The statistical data is given in table 5.

Metric	Count (Transport Domain)
Class count	3678
Object Property count	87
Data Property count	8
SubClassOf axioms count	8045
SubObjectPropertyOf axioms count	76
Individual count	264
Annotation Assertion axioms count	1918
Class Assertion axioms count	258
Number of Assertions	1000 000

Table 5. Statistics of implementation of PurposeNet in OWL

4 Comparison with Other Ontologies

We evaluate PurposeNet from two perspectives:

1. Quality Evaluation in terms of various metrics as tabulated in table 6;
2. Estimation of how well the ontology represents the given search terms in the context of ontology search engine.

4.1 Metric based Comparison

Three popular ontologies were selected for a metric-based comparison with PurposeNet to evaluate its quality. The three ontologies selected are – the general Semantic Web Technology Evaluation Ontology (SWETO) (Meza et. al, 2007), Glycomics Ontology (GlycO) (Satya et. Al, 2005), and, TAP (Guha and McCool, 2003).The results shows that PurposeNet scores much higher than all the other ontology in terms of Class Importance (which determines the importance of a class by the ratio of number of instances connected to the subtree attached to a class Ci in comparison to the total number of instances (I) in the ontology, showing how many classes play a central role compared to other classes). The completeness check (for populating relations, showing the percentage of relation slots filled in by values, thereby determining how well the ontology can be utilized) yielded incompleteness for 7 of the 443 classes defined in the Car subtree.

SNo	Metric	SWETO	TAP	GlycO	PurposeNet
1	Classes	44	6,959	361	3678
2	Relations	101	25	85,637	95
3	Instances	813,217	85,637	660	264
4	Schema Relationship Richness	NA	NA	NA	0.185
5	Schema Inheritance Richness	NA	NA	NA	1.68
6	Schema Attribute Richness	NA	NA	NA	44
7	Class Richness	59.1	0.24	48.1	0.029
8	Class Connectivity	8	6	10	4
9	Class Importance (max. value)	59	31	18	100
10	Cohesion	NA	NA	NA	881
11	Class Relationship Richness (max. value)	NA	NA	NA	100

Table 6. Comparative representation of various ontology metrics

4.2 Comparative rank scores of PurposeNet and akt ontology for browser-retrieval

The efficiency of an ontology can also be determined based on the rank search engines on the web gives. Browser-wise, ontologies are usually ranked based on three criteria – user popularity, evaluation tests and structural criteria (Gangemi, 2006). An ontology may be ranked structurally based on CMM, DEM and SSM (HarithAlani and Christopher Brewster, 2006). We have used the reference ontology of the akt (Advanced Knowledge Technologies) project on extraction and use of knowledge (Motta, 2001). The observations with respect to the various ranking measures in PurposeNet in comparison to the best ontology (ranked 1) outcomes obtained by Alaniet. al. (2005) with respect to the akt reference ontology is tabulated in table 7 below. It is observed that the akt ontology performed better with respect to the CMM (Class Match Measurement, the number of concepts in the ontology that either match (M) or contain the search term (C), that determines how many search terms exactly match with terms in our ontology, that presents the certain degree of detail in the representation of the knowledge concerning that concept) as well as DEM scores(Density, the number of superclasses (U), subclasses (S), attributes (A) and siblings (I) associated with the individual concepts in the ontology), whereas, PurposeNet had a better SSM score(Semantic Similarity, how close related terms are placed in the ontology, where, ontologies that position concepts further away from each other are less likely to represent the knowledge in a coherent and compact manner. It is measured by the path

distance between the two different concepts in question), favoring its faster representation on Swoogle.

Ontology	CMM	DEM	SSM	Total Score
PurposeNet	0.786	0.589	0.413	0.596
akt reference ontology	0.833	0.632	0.250	0.571

Table 7. Comparative rank Scores of PurposeNet ontology and akt ontology

5 Purpose Detection and Extraction

The method of knowledge discovery by manual extraction of data and manual building of PurposeNet ontology is quite exhaustive as several experts are required to put in hours of browsing to find the data corresponding to the concerned features and to incorporate it. This also leads to a slow progress in the creation of a knowledge base that was supposed to finally have a size of a million artifacts. We follow a two-step process for the extraction of data from the web. The first task is to find an appropriate method to detect the presence or absence of a relation. The second step would be to extract the relation from the text that is known to contain the semantic relation. This methodology has been applied on the purpose relation as a case study for generalization across all other relations in PurposeNet.

5.1 Purpose Detection

Sentences containing particular relations have specific structure(s) in terms of a key word or words in a particular order. We select WordNet as the corpus for our work. The principle behind the selection of the WordNet as the corpus is the observation that 70% of the WordNet corpus contains purpose data. We perform automatic detection by transforming the problem of relation detection to a binary classification problem. There are many supervised as well as unsupervised methods of classification that have been graded equally well in other domains. Some of these are the Typed Dependency Parse (Catherine et. Al, 2006), Decision tree forest (http://www.dtreg.com/treeforest.htm), the Naïve Bayes method (Bayes et. Al, 1763), the kernel based Neural Network approach and the more popular Support Vector Machine (Vapnik et. al, 1995) based approach. A comparative study of these various methods of detection of purpose data in table 8 shows that the typed dependency and simple decision trees method of detection gives maximum precision over others. A comparison of the various recall values shows that the typed dependency method has the highest recall. Hence, we suggest the typed dependency method as the most favorable among all methods of purpose detection.

Sno	Method	Precision	Recall	F-Measure
1	Typed dependency	0.84	0.68	0.751
2	Simple Decision Tree	0.83	.67	.74
3	Decision Tree Forest	0.679	0.644	.661
4	Bagging	.755	.619	.68
5	Naïve Bayes	.7	.638	.668
6	Bayes Net	.699	.639	.668
7	RBF Neural Network	.679	.595	.634
8	SVM	.694	.639	.665

Table 8. Comparison of efficiencies of various automatic purpose detection methods

5.2 Purpose Extraction

Our target is to extract the artifact whose purpose is known to be available in text. This section explains the three methods used for extraction of purpose from text: a. Clue Based Extraction, b. Extraction using Typed Dependency Parse and c. Extraction using Surface Text Pattern.

Method	Precision for extraction of (artifact, action) pair given purpose-containing sentences
Purpose clues	69
Surface Text Patterns	88
Typed dependency Parse	98.1

Table 9. Comparativeperformancemeasuresof various purpose extraction methods

Table 9 shows a comparison of the performance of the three methods. The results show that Typed dependency method performs well in extraction of (artifact, purpose) pair. Surface Text Patterns perform well too, considering that the entire web is its corpus, vis-a-vis the other two methods which used offline corpora.

6 Applications

PurposeNet has a number of applications in various reasoning tasks, including Question Answering (QA), provision of online help in web pages, aiding expert systems and broadly in Natural Language Understanding. We describe an application that we have built to evaluate our ontology.

6.1 Domain Specific Question Answering

In this application, a passage is given as input to the automated QA system and the output to a set of questions is obtained. The same task is given to an average car user and the two outputs are compared.

Design.
We have built four alternative modules and each module uses a different resource for producing the answer. Module 1 uses only the passage from where the answer is to be retrieved. Module 2 uses passage and script; Module 3 uses passage and PurposeNet and Module 4 uses passage, script and PurposeNet. We have used a racing car text to test the modules. We have developed a script for racing car. A script (Schank, 1974) is a structure that prescribes a set of circumstances which could be expected to follow on from one another. PurposeNet contains information which is true for an artifact in all circumstances and a script is a structure that prescribes a set of circumstances which could be expected to follow on from one another. It is similar to a thought sequence or a chain of situations which could be anticipated. The components of the script for the text are:

1. Entry Conditions – the conditions that must be satisfied before events in the script can occur.
2. Results – Conditions that will be true after events in script occur.
3. Props – Slots representing objects involved in events.
4. Roles – Persons involved in the events.
5. Track – Variations on the script. Different tracks may share components of the same script.
6. Scenes – The sequence of events that occur. Events are represented in conceptual dependency form.

The theme of car racing can be segmented into 5 scenes: 1. Arranging track; 2. Prepare for the race; 3. The race; 4. The finish; 5. The victory lap.

Script:Car Race	Track: American Car Race – A Win
Props: R = Race Car T = Race Track F = Checkered Flag G = Shot gun	*Roles:* D = Car Driver S = Spectator Q = Pit team O = Organizer

<table>
<tr><td>P = Petrol
L = Finish Line</td><td></td></tr>
<tr><td>Entry Conditions:

• T exists
• R exists
• D exists</td><td>Results:

• D has more money.
• D has won the race.
• R has less P.</td></tr>
<tr><td>Scene1: Arranging the track

• O sprinkles T
• O grinds T
• (go to scene 2)</td><td>Scene 2 : Prepare for Race

• O checks T
• O signals R line-up
• D line-up R
• D test-drive R
• O signals start race with G
• (go to scene 3)</td></tr>
<tr><td>Scene 3: Race

• D accelerates R
• D steers R
• (go to scene 4)</td><td>Scene 4: Finish Race

• D crosses L.
• (go to scene 5)</td></tr>
<tr><td colspan="2">Scene 5: Victory Lap

• D gets F.
• D waves F.
• D drives on T.</td></tr>
</table>

Table 10. A simplified racing script

The complete Script could be described in Figure above.

Result.

Experiments were conducted on answering questions where both the passage and the questions were given as input to each of the 4 modules and compared with theoutput of human users. The results show that the comprehension passage alone yielded 6% of the answers. These were Queries that were directly related to the story in the Comprehension passage, such as –*Did the drivers test-drive?* 10% of the queries related to Car race are answered by PurposeNet alone. These were technical Queries related to Cars such as – *How did the pit Team repair Clint's car tyre?* 27 % of the queries are answered using Scripts alone. These pertained to the sequence of events in a stereotypical Car race, such as – *What is the connection between waving the checkered-flag and the victory-lap?*

SNo	Resource used to obtain Answer	No. of Queries correctly replied (/30)	Efficiency (/30) in %	% of answers given using this resource
1	Comprehension Passage	2	6	12.5
2	Script	8	27	50
3	PurposeNet	6	20	37.5
4	PurposeNet + Script	14 + 3	57	89
Total		19	—	—

Table 11. Comparative results of Queries answered by AOM Script Applier using various resources

7 Conclusion and Future Work

The paper presents the conceptual base, architecture and implementation of a semantic knowledgebase called PurposeNet with an evaluation performed on it comparing it with some other available knowledgebase. Building an exhaustive knowledgebase is a laborious and intense task, it needs human expertise and it needs good web data processing tools so that information from the web can be easily extracted in order to build the knowledgebase semi-automatically. In order to maintain the quality of the resource, we have, till now, manually created the knowledgebase. Nevertheless, we understand that creating such huge resource completely in manual mode would be a time-consuming work. We have noticed that artifact related information which is useful for our knowledgebase is available in various resources such as WordNet, Wikipedia and other web corpora. We have conducted a few experiments on detecting and extracting purpose of artifacts from web corpus and reported the result in this paper. Experimental results in domain-specific question-answering have produced promising results.

References

1. Alani, H., Brewster, C.: Ontology ranking based on the analysis of concept structures. In: Proceedings of the 3rd international conference on Knowledge capture (K-CAP '05). ACM, New York, NY, USA, 51-58 (2005)
2. Alani, H., Brewster, C.:Metrics for ranking ontologies. In: *WWW2006*, May 22–26, 2006, Edinburgh, UK.Harry D Patton. Physiology of Smell and Taste: *Annual Review of Physiology* 12. pp 469–484 (2006)
3. Aleman-Meza, B., Halaschek, C., Sheth A., Arpinar, I. B., Sannapareddy, G.: SWETO: Large-Scale Semantic Web Test-bed. In: Proceedings of the 16th SEKE 2004: Workshop on Ontology in Action, Banff, Canada, June 21-24, 2004, pp.490-493 (2004)
4. Bharati, A., Chaitanya, V., Sangal, R.: Natural Language Processing: A Paninian Perspective", Prentice-Hall of India, New Delhi (1995) (Download:http://ltrc.iiit.ac.in/downloads/nlpbook/nlp-panini.pdf).

5. Bharati, A., Nawathe, S.A.,Chaitanya, V., Sangal, R.: A New Inference Procedure for Conceptual Graphs. In: Proc. of 4th University of New Brunswick Artificial Intelligence Symposium (1991)
6. Cowell, E. B., Gough, A. E.: The Sarva-Darsana-Samgraha or Review of the Different Systems of Hindu Philosophy: Trubner's Oriental Series, Taylor & Francis (2001)
7. Lenat, D. B.: CYC: a large-scale investment in knowledge infrastructure, Communications of the ACM, v.38 n.11, p.33-38 (1995)
8. Gangemi, A., Catenacci, C., Ciaramita, M., Lehmann, J.: Modelling Ontology Evaluation and Validation. In: Proceedings of the 2006 European Semantic Web Conference (2006)
9. Devi, G.: Padartha Vijnana made easy. Chaukhamba Sanskrit Pratishthan, Delhi (2007)
10. Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J.: WordNet: An online lexical database. International Journal of Lexicography, 3(4) (1990)
11. Iśvarakṛṣṇa, Sāṁkhyakārikā with SankaraMisra's commentary Sāṁkhyatattvakaumudi , Edited and translated into Hindi by Nigam Sharma. Varanasi: Parimal Prakashan (2007)
12. Mayee, P.K., Sangal, R., Paul, S.: Action Semantics in PurposeNet. In: Proceedings of 2011 World Congress on Information and Communication Technologies. IEEE WICT'11. pp. 1299-1304 (2011)
13. Kulkarni, Amba P, Navya-Nyaya and Logic, MTech Thesis, IIT Kanpur (1994)
14. Liu, H., Singh, P.: ConceptNet: A Practical Commonsense Reasoning Toolkit. BT Technology Journal, Volume 22. Kluwer Academic Publishers (2004)
15. Nagaraj, A, Manav Vyavhar Darshan, Jivan Vidya Prakashan, Amarkantak, 2003.
16. Praśastapāda, Padārthadharmasamgraha with Sridhara's commentary Nyāyakandali, edited and translated into Hindi by Srī Durgadhara JhaSharma. Vārānasī: Sampurnananda Sanskrita University (1997)
17. Ram Sunder Rao . M. Ayurveda Padardha Vijnana (2003)
18. Sangal, R., Chaitanya, V.: An Intermediate Language for Machine Translation: An Approach based on Sanskrit Using Conceptual Graph Notation, Computer Science and Informatics Journal, Computer Society of India, 17, 1, pp. 9-21 (1987)
19. Singh, N.: Comprehensive Schema of Entities: Vaiśeṣika Category System," Science Philosophy Interface, Vol. 5, No. 2, pp 1-54 (2001)
20. Sowa, J. F.: Conceptual Structures: Information Processing in Mind and Machine, Addison-Wesley, Reading, MA (1984)
21. Sowa, J. F.: The Challenge of Knowledge Soup. In: J. Ramadas & S. Chunawala, Research Trends in Science, Technology, and Mathematics Education, Homi Bhabha Centre, Mumbai, pp. 55-90 (2005)
22. Sowa, John F.: Relating diagrams to logic. In: G. Mineau, B. Moulin, & J. F. Sowa eds., Conceptual Graphs for Knowledge Representation, Springer-Verlag, New York (1993)
23. Tartir, S., Arpinar, I. B., Moore, M., Sheth, A. P., Aleman-Meza, B. OntoQA: Metric-based ontology quality analysis. In *Proceedings of the IEEE Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge Sources (ICDM'05), Boston, MA.* (2005)
24. Varma, V.: Building Large Scale Ontology Networks. Language Engineering Conference (LEC'02) December 13 - 15, 2002. India, p. 121(2002)

Appendix

SNo	Description Feature	Definition	Values

1	Color	The property possessed by an object of producing different sensations on the eye as a result of the way it reflects or emits light	Red, Blue, Green, Yellow, Cyan, Indigo, Orange, Pink, Black, White, Any.
2	Constitution	The material with which an artifact is made up of	Metal, Rubber, Wood, Foam, Plastic, Glass, etc.
3	Fluidity	The physical property of a substance that enables it to flow	Fluid, Nonfluid
4	Heaviness	The comparative weight of an artifact	Heavy, Light, Moderate Weight
5	Inertness	The reactivity of an artifact with the substances around it	Inert, Alkaline, Acidic
6	Mobility	The movement of an artifact during the performance of its target task	Mobile, Immobile
7	Oiliness	The presence of oil on the surface of the artifact	Oily, NonOily
8	Position	The position of an artifact vis-à-vis the artifact it is embedded in	Above, Below, Inside, Left_Of, Right_Of, In_Front_Of, Behind
9	Shape	The external appearance of an artifact	Cubical, Spherical, Circular, Oval, Triangular, Aero, any
10	Size	The amount of space occupied by the artifact	Microscopic, very small, small, medium, large, any
11	Sliminess	The sticky, slippery property of an artifact	Slimy, Nonslimy
12	Smell	The property of an artifact that is sensed by the nose	No odour, Weak, Very Weak, Strong, Intolerable
13	Smoothness	The property of having a surface free from projections or irregularities	Smooth, Rough, Sharp, etc.
14	Softness	The property wherein the artifact gets deformed on application of pressure	Soft, hard
15	Sound	Mechanical vibrations emitted by artifacts when they function	Silent, whisper, bearable_sound, unbearable_sound
16	Stability	Indicates whether the given artifact remains as it is or disintegrates into the environ-	Stable, Unstable

		ment	
17	State	The physical state in which the artifact usually exists	Solid, Liquid, Gas
18	Subtleness	Indicates whether an artifact is so slight that it is difficult to perceive	Subtle, Nonsubtle
19	Taste	Indicates the property of an artifact that is perceived by the tongue	Sweet, Sour, Bitter, Umami, Salty
20	Temperature	Indicates the temperature at which the artifact usually exists	Hot, Cold, Warm, Normal,, Cool
21	Transparency	The property of the surface of an artifact that allows a human to see through it	Transparent, Opaque, Semi-transparent
22	Std. Capacity	Maximum weight that this artifact can hold	….kgs, ….lbs,…ltrs
23	Std. Magnitude	Standard dimensions of the artifact	….metres
24	Std. Weight	Weight of this artifact	….kgs, ….lbs

Static and Dynamic Knowledge Modeling in Geotectonics

Vladimir Anokhin[1] and Biju Longhinos[2]

[1]All-Russia Gramberg Research Institute for Geology and Mineral Resources of the Ocean (VNIIOkeangeologia), Angliisky Prospect, 1, Saint Petersburg 190121 Russian Federation
vladanokhin@yandex.ru

[2] Department of Geology, University College, Trivandrum City, 695 034 India
biju.longhinos@gmail.com

Abstract. Geotectonics, being one of the main geological disciplines, encounters conceptual difficulties that likely can be resolved by application of methods of knowledge engineering. However, a strategy of their application is needed. The role of ontologies in the knowledge-engineering process is to facilitate the construction of a domain model. This model can be either static, i.e., address only the observed geological structures and landforms, or dynamic, accounting for processes that operate in and below the earthcrust. Both types of model are required to overcome the conceptual problems of geotectonics, but while the former is more or less present in the literature, the latter represents complete terra incognita. Meanwhile, exactly the dynamic knowledge modeling is the most important for a field like geotectonics.

Keywords: Plate tectonics, structural geology, knowledge engineering, ontology

1 Introduction

The term “tectonics” originates from a Greek word, “tekton”, which literally means builder. Later this word acquired a wider meaning that included the whole process of creation of something, including such connotations as techniques of construction, properties of the material and principle of creation, or architecture (Laugier’s (1753), Botticher (1852), Semper (1951) and Liu and Lim (2009)). In the Earth sciences, this word is known at least since 1894, when it was said at the 6th International Geological Congress, Switzerland, to describe the mammoth architecture of the Alps and Jura Mountains (Franks and Trumpy, 2005). Since that, the tectonics began to form as a subdiscipline of geology and was defined as the branch of geology that deals with the architecture, or structure, of the outer part of the solid Earth. The same time, the account for regional structural or deformation features and the study of their interrelationship, origin and evolution was referred to another subdiscipline called structural geology. The distinction between the two is often blurred, especially at regional and local scales, as both describe the principles and mechanisms of rock dislocation and

deformation. To handle the ambiguities, a few terminologies were added, e.g., geotectonics for the study of tectonic features in regional scale, global tectonics, for research of tectonic processes related to very large-scale movement of material within the Earth, megatectonics, a tectonics of very large structural features of the world with respect to time (now rarely used).

It is accepted in knowledge engineering that a model in particular domain is built in process of human-computer (or expert – knowledge engineer) interaction and thus largely reflects the way of thinking of the expert. Relying on multiple experts decreases the "human component" but still keeps the record of personal experiences. Here the emphasis is somewhat different; it is not intended to create a cognitive model, i.e. to simulate the cognitive process of expert. Instead, the challenge is to create a model that represents "bare knowledge" and, as such, offers as much bias-free results as possible. While the expert may consciously articulate some parts of his or her knowledge, he or she will not be aware of a significant part of this knowledge since it is hidden in his or her skills. This knowledge is not directly accessible, but has to be built up and structured during the knowledge-acquisition phase. Therefore, at some point (known to or felt by knowledge engineer) the acquisition of knowledge from particular experts (or texts) should be replaced by building a model, desirably as bias-free as possible. Certainly, any model is only an approximation of the reality. Apparently as well, the modeling process is infinite. However, in every knowledge engineering process the stages of knowledge acquisition and model construction should be, from one side, clearly divided, and from the other, tightly interrelated. In our opinion, the best connection between them could be ontology of considered domain of tectonic knowledge. Creation of ontology or relation of extracted knowledge to some pre-existing ontology should be the result of knowledge acquisition and starting point for knowledge modeling.

Ontology provides vocabulary of terms and relations to a model. The closer it is to the domain of interest, the better the model will be. For instance, if ontology perfectly suits the domain, then a domain model in some cases can be obtained just by filling the ontology classes with instances. However, this rarely happens, first, because the nature of ontology is to be generic, while domains of interest usually occur at intersection or as particular cases of such generic domains, and then, because only static model, assuming that modeled environment does not change, can be obtained right from ontology (see below). Also, ontology helps avoid mixture and overlap of meanings and figure out groundless meanings. For example, geologists often use 'subsidence' and 'uplift' to indicate crustal movements against sea level, however, ignoring the fact that the concept "sea level" is related to other concepts which indicate "exterior" phenomena (e.g., river flow discharge or precipitation from atmosphere) that may change simultaneously with crustal movements (i.e., there will be nodes in ontology denoting these exterior phenomena and nodes denoting blocks of the earth crust, and both types of nodes will be bound with the third type, indicating the periods of time, by similar relation, say, "change" or "vary within").

One can evidently see a two-tier division in modeling of tectonics and related disciplines, (i) modeling of the morphologic features, or "anatomy", of the lithosphere,

its interior and surface, and (ii) modeling of the processes that govern the anatomy, i.e., the "physiology" of the lithosphere.

The case (i) implies static entities, like the shape and size of landforms studied by the morphological subdiscipline of geomorphology (when mathematically formalized, this subdiscipline is known as morphometry) or anatomy of geological structures (studied by structural geology that performs description of form, arrangement, representation and analysis of structures that are seen in rocks). It is noteworthy that the "anatomy" of the surface and that of the interior need not to be corresponding each other. Thus, hills may well correspond to synclines and vice versa. Sometimes, if the data are accurately presented, a detailed description may bring an illusion that the static part of scientific research gives full explanation to the phenomenon under study. Still it lacks understanding of the same phenomenon across time and under different parameters.

In case (ii), the processes (i.e., dynamic entities) that govern the anatomy are the focus of study. The dynamic entities change in time and space under some external or internal conditions. Nonetheless, unlike structural geology, this sub-discipline of tectonics has no specific term, though may more or less pass under the term geodynamics. Still, geodynamics is commonly meant to deal specifically with the forces and processes of the interior of the Earth. Therefore, to avoid ambiguity, in this paper the following terms are suggested, morphologic tectonics and dynamical tectonics. Such division is natural for many sciences, e.g., anatomy and physiology (of plants, animals and men), planetary science and cosmogony. "Static" (classification) branches are clearly seen in history, while the main body of its knowledge is certainly "dynamic". In general, one may say, on one hand, that "static" subdisciplines address the composition and structure of systems, and "dynamic", the dynamics, function and evolution of the same systems, in terms of Bogdanov (1926) later replicated by Von Bertalanfi (1968). On the other hand, however, this is fully compliant with the division of knowledge in knowledge engineering into static and dynamic suggested by Pshenichny and Mouromtsev (2013) and earlier formulated classification of methods of knowledge engineering into object-based and event-based, correspondingly (Pshenichny and Kanzheleva, 2011).

2 Purpose and Tasks

Geotectonics encounters conceptual difficulties from perceptional conflicts out of variant interpretation of same observation. The dilemma has to be resolved to bring forth unified scientific approach to earth system understanding. This paper considers the applicability and usefulness of knowledge engineering methods in the study of tectonics. For this, it explores the application of knowledge engineering (i) in morphological tectonics (structural geology) and (ii) in dynamical tectonics. Its main mission is to pave the way to future research in bringing a unified ontology which caters dynamic models in geotectonics as well as in other branches of geology.

3 Knowledge engineering in morphological tectonics

Recent studies revealed a variety of perspectives to deal with object based-methods of knowledge engineering in morphological tectonics. Zong et al. (2009) suggested class hierarchy of geological structures (Fig.1). Similar hierarchies and ontologies exist in other earth-scientific domains, on which the dynamical tectonics is based (Ma, 1980; McGuinness et al., 2007; Sinha et al., 2008, and others). These ontologies scrutinize the field of knowledge and make it computer-understandable. The same time, they do not allow to evaluate how trustful regional data are and to what extent their subjectivity is due to the method of study and to what, due to the scientist's preoccupation.

Poole et al. (2008) suggest an approach that marries ontologies and Bayesian probabilistic computation as a possible solution. Here, the structure of probabilistic theories does not necessarily follow the structure of the ontology. For example, an ontology of lung cancer should specify what lung cancer is, but whether someone has lung cancer depends on many factors of the particular case and not just on other parts of ontologies (e.g., whether they have other cancers and their work history that includes when they worked in bars that allowed smoking). As another example, the probability that a room will be used for living depends not just on properties of that room, but on the properties of other rooms in an apartment. Similarly, in geological parlance, it is difficult to bring interpretation directly from the geological data. For instance, in geological mapping, geologist often tends to see what he wants to see, sometimes departing rather far from the facts – e.g., he "sees" faults which unlikely can be seen, finds stress deformations where an evidence of strain exists, traces rock block displacement in an opposite direction and so forth. The decisions made by geologist are often intuitional. It is observed that the instrumental data, geophysical and others, are being treated very broadly, often solely not to undermine the theory that the geoscientist ''believes'' in. Now adding the probability distributions to the classes of ontology, which describes tectonic study, as proposed by Poole et al. (2008) may give a tool to show how probable is the suggested interpretation of given data. However, the result would not solve the remaining puzzle – the evaluation of the theory itself. An attempt to resolve the problem is addressed below, considering all special cases present in tectonics.

4 Knowledge engineering in dynamical tectonics

All existing theories in tectonics are genetic, that is, they not only involve description of products (usually done within the realm of morphological tectonics) but also involve the description and interpretation of processes. For example, the great mountain arc of Himalayas is not described as a static feature; instead, in tectonics it is considered as a product of ongoing phenomenon of uplift, run either by gravity mechanics (principles of heat engine) or by quantum mechanics (principles of stress engine) (Tassos, 1998). It stresses the claim of Pshenichny and Mouromtsev (2013) that tectonic theories entirely lie in the realm of dynamic knowledge.

Fig. 1. Class hierarchy of Geological Structure. PSS - pressure solution seam and DB - deformation band (Zhong et al., 2009).

During last two centuries, the Earth science saw the rise and fall of many hypotheses that intended to look into the dynamical tectonism. They all can be considered as modern way of looking into the planet Earth and its evolution through the 'absolute' geological time (a concept introduced by Patterson (1953) and Houtermans (1953b). Most of the hypotheses in dynamical tectonics that has been debated fall into one of three classes assuming one of the following states of the Earth, the "contraction" (Beaumont's mountain formation model, the Dana - Hall model, the Suess model, the Barrell model), "expansion" (the Egyed-Jordan model, the Vogel model, the Carey model) and "steady state" (the Hayford-Bowie model, the Kreichgauer model, the Wegner – du Toit model, the Vine-Mathews-Morgan-Wilson model). There are a few more models like plume tectonics, surge tectonics, vortex tectonics, Belousov and Kosygin concepts (see, e.g., Kosygin, 1983), pulsating Earth concepts (Milanovsky, 1995) and the youngest hypothesis, namely the global wrench tectonics (Storetvedt, 2003) are not fit into the above three tier division, which is based on the radius of earth across time. Among the theories listed first, the geosyncline theory (Dana-Hall model) is existing for more than 100 years, though a large space is occupied by plate tectonics model (the Vine-Mathews- Morgan-Wilson model) since 1960s (see, e.g., Morgan,1971).

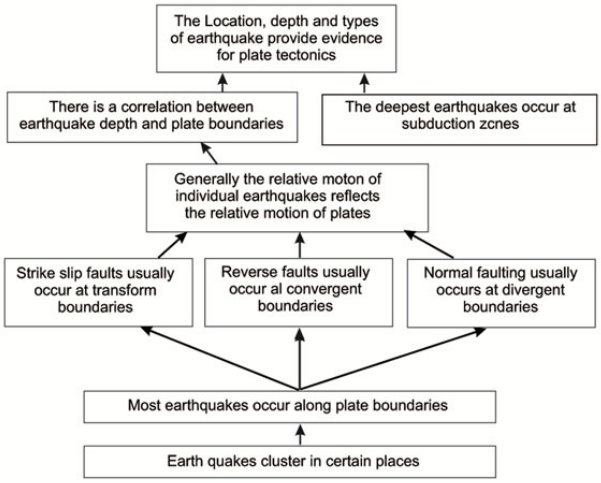
Despite the acceptance or rejection of a model, each of them contains facts which are evident – and each gives sufficient explanation only to a part of such facts. For example, the contraction tectonic school easily interprets tilted strata and mammoth relief features on the globe, but seldom looks at the jig-saw puzzle fit of continents across oceans (the Atlantic case). The hypothesis does not give an apt account of the "stripped pattern" of magnetic anomalies in north Atlantic ridge sector. Similarly, the plate tectonics and the expansion tectonics logically reason the "stripped pattern" of magnetic anomalies and the very existence of middle oceanic ridge structures, but keep silent about the trans-oceanic submerged bridges (having continental characteristics) connecting continents across oceans (Storetveld and Longhinos, 2011; Longhinos, 2012). The coincidence of the morphotectonic features and the subsurface geophysical characteristics across the north-south transect of Australia is interpreted as a deep mantle inflow channel, between Banda Strait (channel outlet) and the Australia-Antarctica Discordance (channel inlet) by the surge and vortex tectonic schools (Leybourne and Adams, 2008). On the contrary, the plate tectonics hardly foresee any Walker type mantle circulation in this tectonically active region (and envisages Hadley type circulation of lithosphere, alone). The tectonic activity in Alpine-Himalayan Belt is another arena of disagreement between hypotheses, where the degree of con-

flict rises with every new piece of data (geosynclinal versus subduction versus wrenching versus vertical uplift). In short, all proposed models in dynamical tectonics cover the truth only partly.

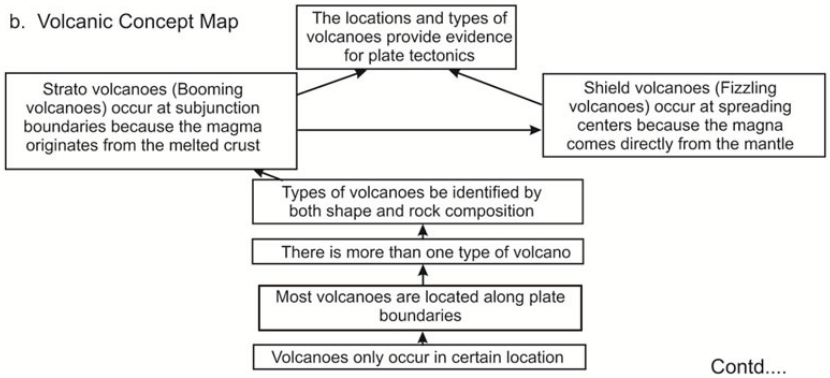
In modern dynamical tectonics, however, only one hypothesis, the plate tectonics, completely dominates. It was in beautiful accordance with the data, mainly geophysical, at the time of its formulation, being the same time amazingly simple and self-consistent. However, many new facts have been reported. In order to fit them, both the theory was modified and interpretations of facts were varied. This has made plate tectonics an object of critique (Meyerhoff and Meyerhoff, 1972; Pratt, 2000, and others). Even though an objective enquiry into the working parameters of plate tectonics has not been attempted so far, so that its application to real data is based on beliefs and assumptions, a graphic conceptualization (Figure 2) somehow substantiating the plate tectonics is presented by Shachter (2007). This is, to the authors' knowledge, one of the very few attempts of "parsing" the structure of this hypothesis proposed so far. Structurally, these graphs have anastomosis patterns (Fig. 2a), multiple paths to a singular node (Fig. 2, a, b, c), ambivalent relations (Fig. 2c) and nested nodes (Fig. 2c), the sense of which is not defined or explicated. Semantically, the graphs do not show definable relationships between the events (i.e., nodes). For instance, looking at Fig. 2a, it looks more or less reasonable the passage from "Earthquakes cluster in certain places" to "Most earthquakes occur along plate boundaries", but it is totally unclear to non-geologist (and to some geologists either!) even from the point of view of natural language why the next step is "Strike-slip faults usually occur at transform boundaries", "Reverse faults usually occur at convergent boundaries" and "Normal faulting faults usually occurs at divergent boundaries". Obviously, an explicit link between "earthquakes" and "faults" should be included in the conceptualization. Also, it is not clear what these diagrams mean to say in general – neither they introduce a theory nor prove it. Perhaps they show the compliance of the theory with considered evidence. Thus, in case of Figs. 2a, b, it clearly shows that compliance is not sufficient, as only "most" earthquakes and volcanoes are considered by the theory, and those minor which occur outside of plate boundaries, are not. However, even sufficient compliance with the evidence is not necessarily an explanation of this evidence, while explanation is exactly the purpose of the theory. Such explanation offered by a theory is not demonstrated by the quoted graphs. Nevertheless, even at this highly informal and superficial level it could be interesting to use such conceptualization for other tectonic theories (plume tectonics, geosyncline theory and others) to show (in)compatibility of theories against similar evidence. Finally, from the point of view of Earth science context, these plots seem to be very general and may appear misleading, as they do not go into necessary detail. E.g., stratovolcanoes and shield volcanoes may be well combined in similar settings and even built on top of one another, despite the enchanting simplicity of their separation in the plot (Fig. 2b). Also, it is not specified what fossils may be really indicative of spatial proximity of areas of their occurrence, while this issue is often debatable in paleontology, and similar fossils are sometimes found in areas which could not be adjacent by the same very theory of plate tectonics (Pratt, 2000).

While the compliance with facts should be likely addressed by the object-based methods as discussed in the previous section, the structure of the theory, as it describes the processes that are believed to operate in the Earth crust and mantle, may be a subject for event-based methods of knowledge engineering. Also, structure of other theories and their compliance with similar facts and with each other should be studied by the whole armory of concept- and event-based methods. These methods are truly new in dynamical tectonics

a. Earthquake Concept Map



b. Volcanic Concept Map



Contd....

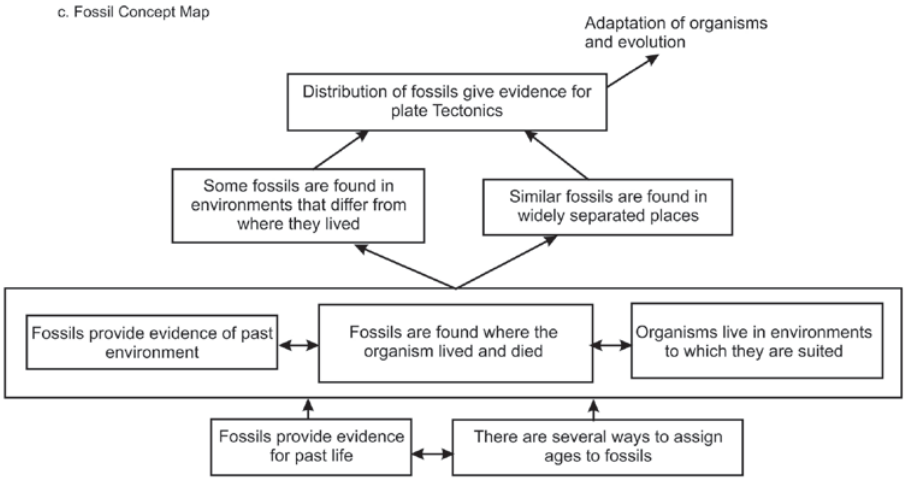


Fig. 2. Concept maps for plate tectonics, a – earthquake evidence-based reasoning map, b – volcanism evidence-based reasoning map, c – fossil evidence-based reasoning map, (Shachter, 2007)

5 Discussion

The modern state of tectonics urges wide application of knowledge engineering approaches mainly to handle psychological and social aspects of this science (traditions, bias, herding and so forth) and to reduce their impact on scientific results. At present, however, these approaches are being dramatically underestimated and underused. Two dimensions of their application are straightforward in tectonics, (i) development of hierarchies and ontologies of geological knowledge used in tectonics and adding probability distributions to determine the probability of interpretation given the data and (ii) plotting the mechanisms suggested by each tectonic theory and corresponding scenarios of Earth evolution. The former corresponds to morphological tectonics, treats knowledge in a static way and can be performed by object-based methods of knowledge engineering. The latter relates to dynamical tectonics, models dynamic knowledge and needs event-based methods of knowledge engineering.

While the hierarchies and ontologies have been abundantly developed and their integration with Bayesian computation based on assumed probability distribution has been discussed (Poole et al., 2008), creation of an event-based framework for geotectonics is a perfect terra incognita. Existing attempts are scarce and methodologically incomplete. Nevertheless, exactly these methods are required to

- Assess whether a theory is self-consistent and how well it covers the domain it pretends to cover (i.e., how well it describes the tectonic processes that lead to the

observed results), identify gaps, uncertainties and ambiguities in a theory, as well as its "protective belts" (Lakatos, 1970) introduced artificially to protect the core of the theory;

- Find out how many alternative theories may describe similar phenomenon;
- If there are a number of theories describing similar phenomenon, determine how well each theory covers the domain of interest;
- Estimate the relevance of contradictions between the theories (which may appear purely verbal)
- Enquire whether the mechanism proposed by a theory must be necessarily global or may operate locally in space or time (for instance, whether the plate tectonics may develop only where the asthenosphere is thick enough to enable the plate motion and whether it may wane and give way to other mechanisms otherwise). and
- Look for compatibility of mechanisms from different theories. For example, spreading of the oceanic floor may appear the case not only a driving force of plate growth in plate tectonics but, without subduction, also a consequence of expansion of the Earth.

6 Conclusions

1. Modern geotectonics requires application of methods and approaches of knowledge engineering.
2. Static knowledge engineering techniques (hierarchies, ontologies and others) work well in structural geology or, broadly speaking, in morphological tectonics.
3. In dynamical tectonics the need for application of knowledge engineering methods is much greater; what is required in this domain is methods of modeling events, states, processes and scenarios, or engineering of dynamic knowledge.
4. These methods have been largely unused in the discussed domain, and up to now, even if used, are applied mainly not to compare theories and develop a self-consistent tectonic body of knowledge but to show the advantages of one given theory.

Acknowledgement. The authors are deeply obliged to Cyril Pshenichny whose enthusiasm largely fueled up the creation of this paper; also, Lev Maslov and Paolo should be thanked for presenting their opinion regarding the work. Jishnu B.K (CUSAT, Cochin) and Santhosh P.R. (Tandem) for helping us with their expertise in graphics.

References

1. Bogdanov, A.: Allgemeine Organisationslehre. Tektologie. Bd. I, II: Berlin, Alexander und Lang (1926).
2. Bötticher, K.: Die Tektonik der Hellenen , Potsdam: Ferdinand Riegel, (1852).
3. Carey, S.V.: The expanding Earth, Elseiver. Amsterdam 548 p. (1976).

4. Franks, S. and Trumpy, R.: The Sixth International Geological Congress: Zürich, 1894, Episodes, Vol. 28, no. 3, pp.187-192 (2005).
5. Kosygin, Yu.A., Tectonics, 2nd edition. Nedra Publishers, Moscow, pp 536 (in Russian) (1983).
6. Lakatos, I.: Falsification and the methodology of scientific research programmes. In Criticism and the Growth of Knowledge, LAKATOS, I., MUSGRAVE, (Eds.) Cambridge University Press (1970).
7. Laugier, M.A.: An Essay on Architecture, translated by Wolfgang and Anni Hermann, Los Angeles (1753) .
8. Liu, Y and, Lim, C.K.: New Tectonics: Towards a New Theory of Digital Architecture : Birkhäuser, 206 p. (2009).
9. Longhinos, B.: The Shetland-Greenland land bridge contradicting Atlantic seafloor spreading, Proceedings of the 34th International Geological Congress, Theme 37, Paper2, Brisbane, Australia (2012).
10. Ma, X.: Ontological spectrum for geological data interoperability, Ph.D thesis, University of Twente, ITC Printing Dept Pub. Netherland (2011).
11. McGuinness, D. L., Sinha, A. K., Fox, P., Raskin, R., Heiken, G., Barnes, C., Wohletz, K., Venezky, D., and Lin, K.: Towards a Reference Volcano Ontology for Semantic Scientific Data Integration. American Geophysical Union, Fall Meeting 2007, abstract #IN42A-03 (2007)
12. Meyerhoff A.A., Meyerhoff H.A. "The New global tectonics": Major Sucosn sistencies // Am. Assoc. Petr. Geol. Bull. Vol. 5, No 2. P. 269-336 (1972).
13. Milanovsky, E.E.: Earth Pulsation. Geotektonika, no. 5, pp 3-24 (in Russian) (1995).
14. Morgan W.J.: Plate motions and deep mantle convection, in: Shagam R., ed., Hess Volume, Geol.Soc. Am., Mem., 132, (1971).
15. Poole, D., Smyth, C. and Sharma, R: Semantic Science: Ontologies, Data and Probabilistic Theorie in Paulo C.G. da Costa, Claudia d'Amato, Nicola Fanizzi, Kathryn B. Laskey, Ken Laskey, Thomas Lukasiewicz, Matthias Nickles, and Mike Pool (Eds.), Uncertainty Reasoning for the Semantic Web I Springer LNAI/LNCS (2008)
16. Pratt, D.: Plate Tectonics: A Paradigm Under Threat. Journal of Scientific Exploration, vol. 14, no. 3, pp. 307-352, (2000)
17. Pshenichny, C.A., and Kanzheleva, O.M.: Theoretical foundations of the event bush meth-od. In Societal Challenges and Geoinformatics, GSA Special Paper 482, Sinha, K, Gunder-sen, L., Jackson, J., and Arctur, D. (Eds.), 139-165.(2011)
18. Pshenichny, C.A., and Mouromtsev, D.I.: Representation of the Event Bush Approach in Terms of Directed Hypergraphs. In: ICCS Proceedings, Mumbai, 2013 (in press) (2013)
19. Sinha, K., Malik, Z., Raskin, R., Barnes, C., Fox, P., McGuinness, D. and Lin, K.: Semantics-based Interoperability Framework for Geosciences. AGU Fall Meeting Abstracts. (2008)
20. Shachter, R. D.: Building with Belief Networks and Influence Diagrams, in Advances in Decision Analysis: From Foundations to Applications, 2007, edited by W. Edwards, J. Ralph F. Miles and D. v. Winterfeldt, Cambridge University Press: 177-201 (2007).

21. Storetvedt, M.K.: Global Wrench Tectonics, Fagbokforlaget, Bergen 397 p. (2003).
22. Storetveldt, M.K. and Longhinos, B.: Evolution of North Atlantic: Paradigm shift in the offing. New Concepts in Global Tectonics Newsletter, No.59, pp 8- 51 (2011).
23. Tassos, S.T.:The Cognitive tools of earth expansion, Proc. Int. Symposium on New Concepts of Global Tectonics, Tsuhuba, Japan pp 188-193 (1998).
24. Von Bertalanfi, L.: General System theory: Foundations, Development, Applications: George Braziller, New York (1968)
25. Zhong J., Aydina A. and McGuinness D. L.: Ontology of fractures. Journal of Structural Geology, Volume 31, Issue 3, March 2009, Pages 251–259 (2009).

A Method for Data Minimization in Personal Information Sharing

Prima Gustiene and Remigijus Gustas

Department of Information Systems, Karlstad University, Sweden
{Prima.Gustiene, Remigijus.Gustas}@kau.se

Abstract. A fundamental privacy principle, which is enforced in many privacy-enhancing technologies, is data minimization, i.e. the amount of personal data that are revealed to others and extend to which they are processed should be minimized. Privacy-enhancing identity management is important for processing personal data, the purpose of which is to protect personal data. This is especially relevant for communication via Internet where users are leaving much personal data. Privacy issues should be embedded into a system's core functionality. Minimization of data should be maintained and controlled throughout the systems lifecycle, from the early stages of system analysis and design to implementation. The primary goal of this paper is to present a conceptual modelling method, including the framework, modelling process and the basic modelling constructs, which enables minimization of data. Data cannot be analysed separately without taken into account the processes that cause the changes of data as well as goals. Analysis of relevant data contributes to the problem of data minimization in privacy-enhancing technologies.

Keywords: conceptual modelling, data analysis and minimization, privacy enhancement, service-oriented modelling method.

1 Introduction

Privacy is an essential and fundamental human right (Schütz & Friedewald, 2011). Growing technological possibilities enable transformation of our society towards a computerised social community highly dependent on information sharing. Information sharing increases privacy problems e.g., unauthorised access to personal data and using it for unexpected purposes. Privacy involves the protection of personal information. Information privacy is one of the aspects of the concept of privacy,

which is related to the person's right to determine what, when and how personal information can be communicated to various recipients (Westin, 1967). The best protection of personal information is when the information is not revealed at all, but this is not possible in this computerised society.

Nowadays business more and more takes place on the Internet. This situation increases problems to secure personal data. The processing of the personal data is usually not transparent for the users, but it can have painful consequences to privacy and security of the users. Sophisticated technologies provide possibilities to trace, store and use non-protected data for different purposes. The question is *what* data and *how much data* should be collected, stored and shared doing business online. The answer could be found in data analysis and design methods, which could help to analyse and minimize the amount of personal data that are revealed to others. The problem today is that the methods are not pragmatic-driven and they have not enough semantic power for analysis of data.

Privacy-enhancing identity management (IDM) systems are designed to enforce legal privacy requirements to guarantee the processing of privacy compliant data (Fischer-Hübner & Hedbom, 2008). IDM provides technical means that enables protection of legal privacy principles. Identity management systems manage different identities of a person, helping in processing their personal data and making data life cycle management more transparent. Difficult problems in the area of personal data processing such as lack of data minimization and data life cycle management require the new research solutions for reducing problems of personal data processing. There is a lack of a method that provides systematic guiding principles for aligning the overall systems design with respect to privacy-enhancement mechanisms. These mechanisms must be analysed in the context of a larger inter-organisational system between service requester and service provider. As every system is unique, privacy issues should be integrated with other types of system requirements. To integrate privacy issues into the whole enterprise development system, it is necessary to consider these issues, as a part of a new system development life cycle, from the early stages of business goals analysis, requirements analysis and design to delivery of the system.

Data is an important asset concerning information privacy. Privacy-enhancing identity management systems require accurate data analysis for processing personal data. Data is created, processed and consumed in various transactions for different operational and analytical purposes. As data is the main element for the management of information privacy, it is critical to identify a minimal amount of data, which are relevant in the specific context or scenario. Interdependencies among models that represent different aspects of the system cannot be analysed in isolation. Business data and business processes should be analysed together. Just having an integrated and systematic modelling method provides possibility to analyse structural, interactive and behavioural aspects of a system together. Such method can be applied for the analysis of different scenarios including such non-functional requirements as information privacy issues. The main goal of this paper is to present a framework of a service-oriented modelling method and a modelling process that could be applied for data analysis and design to solve data minimisation problems.

2 Model Driven Architecture and Modelling Levels

Model Driven Architecture (MDA) (OMG, 2010) is an approach for model-driven engineering of software systems. It provides a set of guidelines for the structuring of specifications, which are expressed as three models: Computation Independent Model (CIM), Platform Independent Model (PIM) and Platform Specific Model. The important feature of MDA is the mapping rules and techniques used to modify one model into another. Still, a question mapping rule automation between CIM and PIM levels is feasible and remains a major research effort. Pragmatic and semantic levels are supporting computation-neutral modelling (CNM) and are similar to CIM at MDA framework (see figure 1). Technology-oriented design is done at a syntactic level. It is implementation specific modelling (ISM), which can be compared with PIM at MDA framework.

Pragmatic specifications aim to provide motivation for conceptual representations of enterprise components at the semantic level that defines business processes across organisational and technical system boundaries. Pragmatic knowledge, expressed in terms of pragmatic entities such as goals, problems and opportunities, provides motivation for conceptual representations of enterprise components, which take part in various business processes of an enterprise. To structure the pragmatic knowledge about business processes (as services) is important, because such knowledge provides motivation for various configurations of service architectures and defines the *'why'* aspect of the problem domain. Pragmatic specifications are necessary for several reasons. They motivate service events and show the guidelines over how pragmatic aspects are mapped to conceptual representations, which define the semantics of business design, including the structural, behavioural and interactive aspects of business processes.

Pragmatic dependencies can be viewed as modelling basis to reason about the intentions of designers related to new solutions. Pragmatic entities such as goals, problems, and opportunities can be related by pragmatic dependencies and analysed in different situations. Any business process functionality can be defined as a service. A service, in different contexts, from pragmatic point of view can be regarded as different pragmatic entities such as a problem, opportunity or a goal. Business activities can be defined in terms of a set of interaction loops between service requester and a service provider, and linked to design goals (Gustas & Gustiene, 2008). Behind every business process is a clear motivation or goal, which can be analysed together with a final process state. The achievement of this state should bring value to customer. Goal hierarchies can help to identify missing processes and data. Goals also provide a basis for reasoning about the semantic incompleteness of system specifications.

Pragmatic specifications aim to provide justification for conceptual representations of data and processes. To understand *how* and *why* technical system components are useful and how they fit into the overall organizational system at least three modelling levels of information system specifications are necessary: pragmatic level, semantic level and syntactic level (Gustas & Gustiene, 2009). According to FRISCO report (Falkenberg et al., 1996) these three levels are of great interest in the context of information systems design as they deal with the usage, meaning, and structures of

system representations. Architectural framework for service-oriented modelling is presented in figure 1 and is explained below.

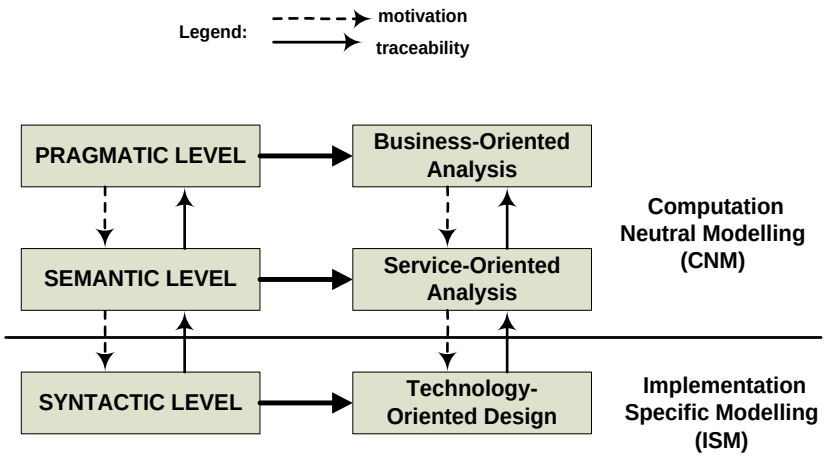


Fig.1. Architectural framework for service-oriented modelling

Pragmatic level is the level where business-oriented analysis is done. Analysis at this level using pragmatic dependencies (Gustiene, 2010) is supposed to drive a system engineering process from business goals to service interactions. Here the decision of which information is necessary and why takes place. *Semantic level* is important for integration of interactive 'where' and 'who', structural 'what' and behavioural 'how' and 'when' aspects (Zachman, 1987) of conceptual representations. Semantic descriptions constrain the implementation specific representations. *Syntactic level* defines the details, which explain the data processing needs for specific application or software component.

The fitness of system specifications between these levels is critical for the success of the final result. Consistency between levels much depends on having appropriate modelling techniques for the refinement of pragmatic entities that justify and represent their structural and dynamic aspects at the semantic level. An integrated modelling way provides possibility to check consistency between levels. It also underlines the possibilities for requirements traceability, which is crucial for verification and validation of system requirements (Maciaszek, 2005). Such three-level architectural framework is the foundation of modelling that helps to provide interplay between business-oriented analysis, service-oriented analysis and technology-oriented design, which are critical for relevant data and process analysis.

3 Perspectives of Service-Oriented Modelling Method

Service-oriented modelling method for information system design (Gustiene,

2010), (Gustas & Gustiene, 2012) is a method based on new principles of service-oriented analysis and design. This method puts into foreground the modelling of interactions (Dietz, 2001) among various enterprise actors. From ontological point of view (Dietz, 2006), every enterprise system could be seen as a composition of different actors that could be viewed as organizational and technical components. The interaction among them is motivated according to the strategic goals of some specific scenario that could be seen as part of some problem domain. The uniqueness of the method is that it is based on the principles of service orientation. Business processes can be analysed as a composition of service interactions. Service-oriented representations are built by conceptualizing interactions among organizational and technical components, which are viewed as various types of enterprise actors. Continuity of interaction loops is the main principle of service orientation. Modelling of interactions between different types of enterprise actors is critical from system analysis and design point of view for several reasons. Explicit modelling of interactions helps to develop an integrated graphical representation of business data and processes. Interaction dependencies among different enterprise actors are important for motivating data transition events and effects (Gustas, 2011). Interaction dependencies provide the possibility to preserve the modularity of crosscutting concerns between different components and to integrate behavioural effects with structural changes in various classes of objects, which represent different data.

The definition of service presented in this method explains the necessary elements and provides with the guidelines how these elements are related to form service architecture. There are two ontological perspectives of communication action (Dietz, 2006), (Gustas & Gustiene, 2009) that lie in the foundation of the main construct for service-oriented modelling: *intersubjective* and *objective.*

The intersubjective perspective defines how actors (service requesters and service providers) are related to each other. This perspective is important as it presents the actors, as independent loosely coupled components, who add value by performing some activities. It signifies certain commitment and responsibilities between enterprise actors (Ferrario & Guarino, 2008).

The *objective* perspective defines how different objects change when the actions during interaction process take place. The objective perspective can be applied to represent the internal behaviour of the objects. It represents data, which is analysed in the context of interaction between organizational and technical system components. Interactions among different actors can be used to manifest object property changes that are results of different actions. Property changes are important for eliciting of semantic meaning of the problem domain. The cohesion of these two perspectives results into a single modelling notation (construct), which allows the integration of static and dynamic aspects of the system, which are important to maintain a holistic representation where external and internal views of service conceptualizations are visualized together.

A starting point of the ontological definition of an enterprise system in the presented service-oriented foundation is quite similar to ontological understanding of system and enterprise as a system. Enterprise system is a composition of the organizational and technical components, which are viewed as various types of enterprise *actors* and which interact as service requesters and service providers. Actors are

subsystems that are represented by individuals, organizations and their divisions or roles, which denote groups of people. Technical actors are subsystems such as machines, software and hardware components, etc. Any two actors can be linked by inheritance, composition, classification or interaction dependencies, which are represented graphically in figure 2.

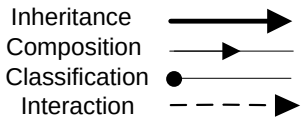


Fig.2. Actor dependencies

Inheritance dependency between actors is used for sharing the static and dynamic similarities. More specific actors inherit the composition and interaction dependencies from more general actors. Dependencies represent additional intrinsic actor interoperation features and structural properties that are prescribed by an enterprise system. *Composition* is a conceptual dependency used to relate a whole to other concepts that are viewed as parts. It is a stricter semantic relation as compared to an aggregation and a composition that is defined in the object-oriented approaches.

Classification link between two actors is used to define their instances. In conceptual modelling, an instance can be viewed as an element of a set that is defined by a concept it belongs to. In the same way as an object can be manipulated by operations, an actor has interaction privileges and responsibilities that are defined by the interaction dependencies.

Interaction dependencies are used to conceptualize services between various enterprise system actors. Since actors are implemented as organizational and technical system components, they can use each other according to prescribed patterns to achieve their goals. Two interaction dependencies into opposite directions between a service requester and service provider define a typical action workflow loop. Interaction flows of a conference management system are represented in figure 3.

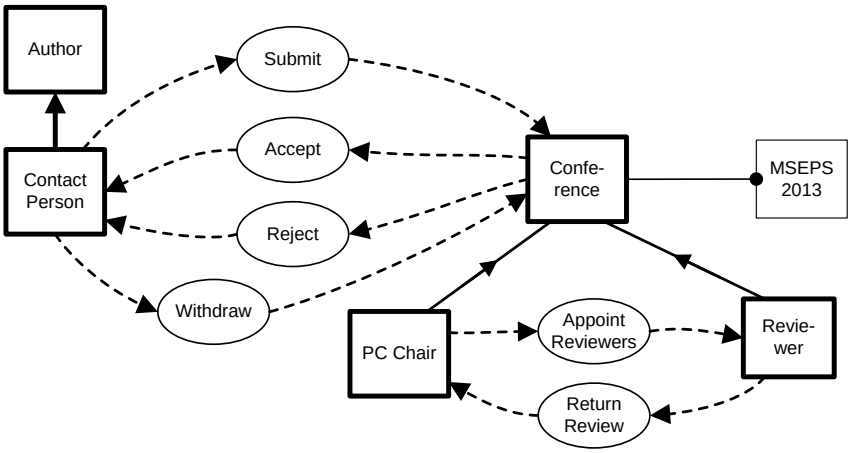


Fig. 3. Example of semantic dependencies between actors

Identification of interaction flows and static dependencies among actors is the first step in the modelling process. An interaction link between two actors indicates that one actor depends on another actor by a specific action. It represents an intersubjective perspective of interaction. For instance, a contact person submits a paper to the conference. He wants this paper to be published. It can be done by accept action, which is initiated by the PC chair. The conference has quite different goals. For instance, conference goals would be to make the submission and reviewing processes as smooth as possible and to be the best conference, i.e. to accept just the best papers.

There are two actors involved in this business process, a contact person who will submit a paper and a conference. A conference is composed of PC Chairs and Reviewers. A conference management system has delegated all major communication through a PC chair, which is a part of a Conference (see composition dependency). PC Chair has a goal to handle reviewing process as good as possible that is to do everything in time. He appoints reviewers and sends the review results to a contact person. The task of reviewers is to reviews the papers. A contact person who is also an author (see inheritance dependency) submits the paper to the conference, by triggering the action Submit. When the person submits the paper, the conference has two possibilities the paper will be accepted, or rejected (see actions Accept and Reject). A Contact Person has also possibility to withdraw the paper (see action Withdraw). When the conference receives the submitted paper, a PC Chair chooses reviewers and sends the paper for review. After reviewing process, a Reviewer returns review to PC Chair (see action Return Review). The results of the review will be sent to contact person (acceptance information or rejection).

4 Modeling Process

The main contribution of this paper is to present the modelling process that consists of five fundamental steps, which support the incremental and systematic service-oriented analysis and design process. An integrated modelling process provides the guidelines for the transition between levels (see figure 1). The requirements traceability is critical for change management, verification and validation processes. A starting point of service-oriented analysis is identification of actor goals and interaction dependencies among service requesters and service providers. The structural aspects of a system are used to represent business data. The behavioural aspects are clarified by defining object transition effects. Without ability to represent noteworthy structural changes, it would be difficult to understand the deep semantics of interactions. Having possibility to reiterate these modelling steps, helps to keep data minimal. The steps of this process are as follows:

1. Identification of the main scenario.

Suppose the conference needs to accept just the best papers and a contact person hopes that his paper to be among those accepted papers. This scenario represents a normal flow of events (Cockburn, 2001). It can be expressed by using two interaction loops, which are represented in figure 3. The first service loop related to paper submission and acceptance. It can be represented as follows:

if Submit(Contact Person ▪▪▪▪▶ PC Chair)

then Accept(PC Chair ▪▪▪▪▶ Contact Person).

The second interaction loop deals with the actions of appointing reviewers and retuning reviews. It is as follows:

if Appoint Reviewers(PC Chair ▪▪▪▪▶ Reviewer)

then Return Review(Reviewer ▪▪▪▪▶ PC Chair).

The interaction flows among actors are graphically illustrated in figure 4.

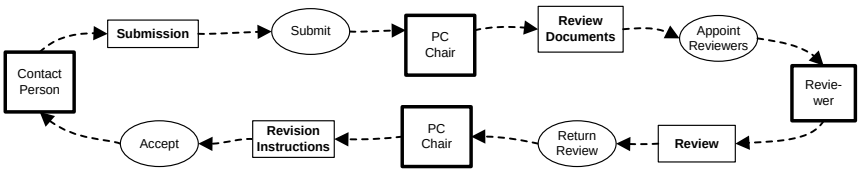


Fig. 4. Main interactions in a conference management system

Interaction flows are the special types of concepts that represent moving flows. In service-oriented modelling method, solid rectangles are used for the denotation of material flows and light boxes show information flows. An action with a missing data or material flow is understood as a decision or control flow. Actions are performed by actors and are represented by ellipses. They are necessary for transferring flows between subsystems, which are represented by various organizational components. Actors are denoted by square rectangles.

2. Definition of actions in terms of transition dependencies.

The internal effects of objects can be expressed by using transition links (⟶) between various classes of objects. There are three fundamental ways for representing object behaviour by using reclassification, creation and termination actions (Gustas and Gustiene, 2012). If termination and creation action is performed at the same time, then it is called a reclassification action. The graphical examples of creation, termination and reclassification are presented in figure 5.

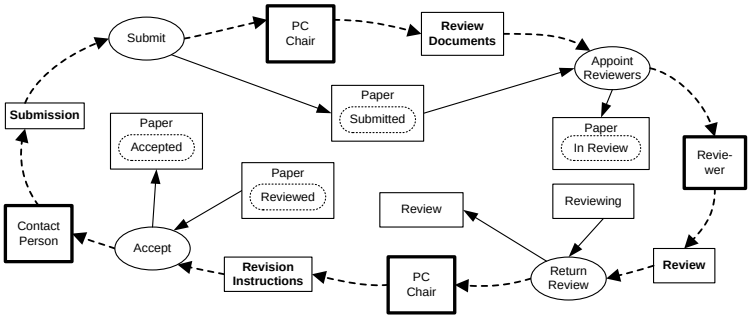


Fig. 5. Main reclassifications in a conference management system

A contact person has a possibility to *submit* a paper. The submission is performed when the Paper[Submitted] object is created. When it is accepted, the responsibility of the conference PC chair is to trigger the *appoint reviewers* action. It is used to send review documents to the reviewers and reclassify Paper[Submitted] to Paper[In review]. Reviewer is obliged to deliver review to PC chair by triggering the *return review* action, which terminates the Reviewing process and created a finalized Review. The PC Chair is authorized to *accept* a reviewed paper by informing a contact person with revision instructions. A Paper[Reviewed] is reclassified to Paper[Accepted] by the Accept action.

3. Identification of a noteworthy semantic difference in every action.

This step is important for identification of attributes, which are affected during object transitions from the pre-condition to post-condition classes. The semantic difference must be defined for every transition dependency by using mandatory attribute links. Various types of attribute dependencies in a conference management system are represented in figure 6. For example, the Accept action changes the state of a Paper object from Reviewed to Accepted. Note that each Paper[Accepted] must be characterized by five properties (Presentation Time, Review, List of Authors, Submission number, Contact Person) and Paper[Reviewed] - by four properties. The noteworthy semantic difference is represented by the complementary attribute Presentation Time.

4. Refactoring.

It is difficult to get initial diagram without inconsistencies and redundancies from the start. Classes and their attributes must be revisited and their semantics examined several times. The refactoring step is necessary to keep conceptual models clean from inconsistent attributes as well as minimize diagrams as much as possible. The refactoring process (Fowler, 1999) does not alter semantics of specification. Refac-

toring is an essential characteristic of good engineering, because it makes necessary structural changes in order to make modelling clean and understandable. The diagram, which illustrates the outcome of this step, is presented in figure 6.

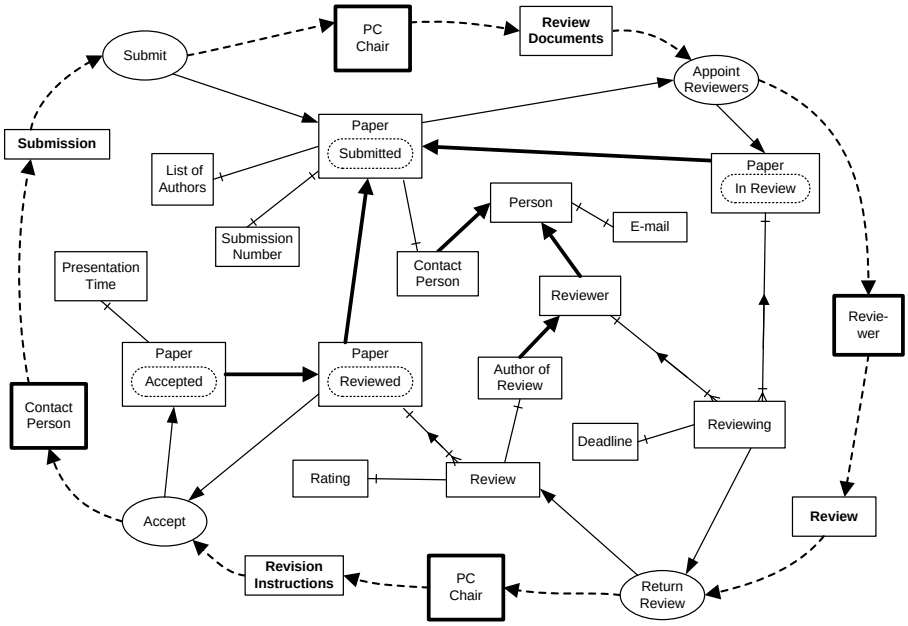


Fig. 6. Interactive, behavioural and structural aspects of a conference management system

Inheritance mechanism allows sharing attributes via generalization/specialization relations. So, inheritance hierarchies can be used to reduce the diagram. For instance, Accepted Paper class in this diagram inherits attributes from Paper in state Reviewed.

5. Describing the alternative scenario.

This step is important, because the modelling process should provide with the possibility to demonstrate available alternatives to the main course of events. Note that the Reject event is an alternative to Accept. Therefore, it should be added at this step (see figure 3).

We have introduced an incremental way of modelling enforcing only minimal data sets, which are represented for various types of actions. Five steps of analysis process are also important for integrity control between static and dynamic aspects in a system. The data necessary in this business process are limited just to most relevant, adequate and not excessive. But the process of data minimisation alone will not solve the problem. Creation of personal data records must be adequate with respect to basic activities in a transaction. It means that analysis of data should be done taking into account the context of business process as well as data handling policy that the service requester and service provider establish in an agreement. How to find out

which data are relevant to specific business process depends much not only on structural analysis, but on interactions and behavioural aspects of a system.

Data life cycles vary in different scenarios. Data life cycle analysis is important to understand and to justify *why* and *when* personal data can be stored in the system and *when* it should be deleted. Creation of personal data records and keeping them in identifiable form not longer than necessary is important principle of privacy enhancement technologies. To manage data life cycle implies analysis of different aspects of data, *what* data is processed, by *whom*, *why*, *where*, and *how*. To get answers to all these questions a holistic integrated representation of a system should be analysed.

5 Conclusion

Problems in the area of personal data processing such as lack of data minimization and data life cycle management require the new research solutions. Privacy issues are always embedded into some organization and are related to business scenarios. It means that these scenarios should be analysed and designed together with functional requirements. As data is the main element concerning the management of information privacy, it is critical to have a way of data analysis for the specific context of business scenario. The advantage of this conceptual modelling method is that it facilitates reasoning about semantic integrity of data. It provides a modelling process and modelling techniques for early requirements analysis, where pragmatic and semantic aspects of different scenarios can be analysed together.

The modelling process supports the traceability between different levels of architectural framework, which is critical for understanding how and why technical system components are useful and how they fit into overall organizational system. Analysis using this method can be applied to solve the problems of data minimization. Applying a method for analysis and design of privacy concerns would contribute with a new knowledge in designing security assurance and privacy-enhancing mechanisms. The method could be applied for diagnosing the *redundant data,* i.e. to distinguish between object properties, which are relevant or not justified with respect to scenarios in the secondary interaction loops. It can be also used to detect the *temporal data* or to distinguish between object properties that can be accessible not longer than necessary. The method is able to justify the *persistent data*, which must be to retained in relation to data transfer scenarios and policies, if various commitments are broken.

References

1. Cockburn, A.: *Writing Effective Use Cases*. Boston: Addison- Wesley (2001)
2. Dietz J. L. G.: Enterprise Ontology: Theory and Methodology, Springer, Berlin (2006)
3. Dietz J. L. G.: DEMO: Towards a Discipline of Organisational Engineering. *European Journal of Operational Research.* 128(2), 351-363 (2001)

4. Falkenberg, E. D., Hesse, W., Lingreen, P., Nilsson, B. E., Oei, J.L.H., Rolland, C., et al.: *A Framework of Information System Concepts* (Report of the IFIP WP 8.1 Task Group Frisco). Leiden; Leiden University (1996)
5. Ferrario, R.. & Guarino, N.: Towards an Ontological Foundation for Service Science. In *Future Internet – FIS2008: The First Internet Symposium, FIS 2008 Vienna, Austria. Revised Selected Papers*, pp. 152-169. Berlin: Springer (2008)
6. Fisher-Hübner, S. & Hedbom, H.: Benefits of privacy-enhancing identity management. Asia-Pacific Business Review, IV (4), 36-52 (2008)
7. Fowler, M.: *Analysis Patterns: Reusable Object Models.* Menlo Park: Addison-Wesley (1997)
8. Gustas, R.: Modeling Approach for Integration and Evolution of Information System Conceptualizations, *International Journal of Information System Modeling and Design,* Vol. 1, Issue No 1, pp.79-108 (2011)
9. Gustas, R. and Gustiene, P.: Conceptual Modeling Method for Separation of Concerns and Integration of Structure and Behavior, *International Journal of Information System Modeling and Design*, Vol. 3, Issue No 1, pp.48-77 (2012)
10. Gustas, R. & Gustiene, P.: A New Method for Conceptual Modelling of Information Systems. *Information Systems Development: Towards a Service Provision Society. Proceedings of the 17th International Conference on Information System Development*, pp.157-165. New York: Springer (2009)
11. Gustas, R. & Gustiene P.: "Pragmatic – Driven Approach for Service-Oriented Analysis and Design", *Information Systems Engineering - from Data Analysis to Process Networks*, IGI Global, USA (2008)
12. Gustiene, P.: Development of a New Service-Oriented Modelling Method for Information Systems Analysis and Design. Doctoral Thesis, Karlstad: Karlstad University Studies, 2010:19 (2010)
13. Maciaszek, L. A.: *Requirements Analysis and System Design,* Addison Wesley (2005).
14. Schütz, P. & Friedewald, M.: Privacy: What Are We Actually Talking About? A Multidisciplinary Approach. In S. Fischer-Hübner, P. Duquenoy, M. Hansen, R. Leenes & Ge Zhang (Eds.) *Privacy and Identity Management for Life.* IFIP AICT 352, pp. 1-14, New York: Springer (2011)
15. OMG.: Model Driven Architecture [Electronic Version]. Retrieved October, 2012, from http://www.omg.org/mda (2010)
16. Zachman, J. A.: A Framework for Information Systems Architecture. *IBM System Journal, 26(3),* 276-292 (1987)
17. Westin, A.F.: *Privacy and Freedom.* Atheneum, New York, NY, USA (1967)

Engineering of Knowledge Structures: Perspectives from Traditional Disciplines and Systems Principles

Doji Samson Lokku, Anuradha Alladi

Tata Consultancy Services, Hyderabad, India
{doji.lokku@tcs.com, alladi.anuradha@tcs.com}

Abstract. A human endeavor can be seen as leveraging certain 'means' in order to accomplish a given purpose. The means could be several, but inevitably knowledge underlies all of them. The proposed paper attempts to present various perspectives that can aid in order to arrive at the 'means' of a knowledge structure.
As per systems methodology, it is the structure and the associated process that when brought together in a given context, give rise to accomplishing a purpose. Accordingly the means could be seen as both the structure and the associated process. Also the concerns that are needed to be taken into consideration while accomplishing a given purpose can be broadly categorized into two, relative to the purpose, in terms of, whether they are *in favor of* or *not in favor of*.
The representation proposed in the paper serves as a guidance while arriving at knowledge structures.

Keywords: Knowledge, Structure, Systems, Purpose, Process, Function, Context

1 Introduction

The epitome of any human endeavor is accomplishment of a stated or intended purpose. Thus a human endeavor can be seen as leveraging certain 'means' in order to accomplish a given purpose. The means could be through technology or knowledge or a human intervention or things such as those, but inevitably knowledge underlies all of them. The proposed paper attempts to present various perspectives that can aid in order to arrive at the 'means'of a knowledge structure. The aim of this paper is to discuss these perspectives with active participation from the audience using familiar illustrations, so as to reinforce our understanding of the topic.

Traditionally engineering is a profession which addresses the concerns of sustenance (while accomplishing a purpose). Accordingly engineering offers sustenance to a means, as the means aids in accomplishing a purpose. On a smaller scope, a purpose can be viewed as a set of functions. As per systems methodology, it is the structure and the associated process that when brought together in a given context, give rise to accomplishing a function. Accordingly the means could be seen as both the structure and the associated process. For example, the 'means' for crossing a river could be a bridge structure.

Principles of systems lend us a handle to capture the concept behind the representation for accomplishing a given purpose. The representation inherits its basis from the principles namely purposefulness, openness, multidimensionality, **counterintuitiveness** and emergent property. The scope of the proposed paper takes into consideration a majority of these principles.

The very many concerns that are needed to be taken into consideration while accomplishing a given purpose can be broadly categorized into two, in relation to the purpose, in terms of whether they are in favor of or not in favor of. Accordingly addressing these varied concerns is expected to result in either accomplishing the purpose to the fullest extent or be able to just cope with, without in any way contributing to the purpose. The former is a fully favorable scenario and the later is a worst case scenario, while accomplishing a given purpose.

For example, a physical structure is expected to withstand seismic or wind forces that come upon it. Where as the sun light that comes into a living room should be maximally utilized towards a healthy living. In other words, the concerns that are in favor lead to thriving and the concerns that are not in favor expects surviving, while accomplishing the purpose. With this line of thought, engineering should ensure that enough sustenance is built into the respective structure & process, both of which are expected to aid in accomplishing a purpose.

This particular representation will find its use in being able to look for the category of concerns that need to be addressed and also to identify the gaps in arriving at the means with respect to the ideal. The currently existing knowledge structures can be seen in the light of this paper and be analyzed for any gaps that exist in them. This representation will also point to the various bodies of knowledge that are multidimensional, using which the varied concerns while accomplishing a given purpose may be addressed.

The representation proposed here for presentation operates at the level of the knowledge that is required for accomplishing a purpose. What are the knowledge structures for ‘Organizing Knowledge’? Familiar human endeavors such as these are attempted for representation, as part of illustration towards the proposed paper.

For the scope of this paper, a human endeavor is viewed as an attempt by people leveraging a certain means, towards accomplishing a given purpose. The purpose could be stated or intended. Also, for the scope of this paper, the purpose which people attempt to accomplish is merely given and as such there is no debate in this paper about the topic of purpose itself, in terms of what is a purpose and why is it a purpose and the related discussion. This particular basic premise about a human endeavor aimed at a purpose is depicted in figure 1.

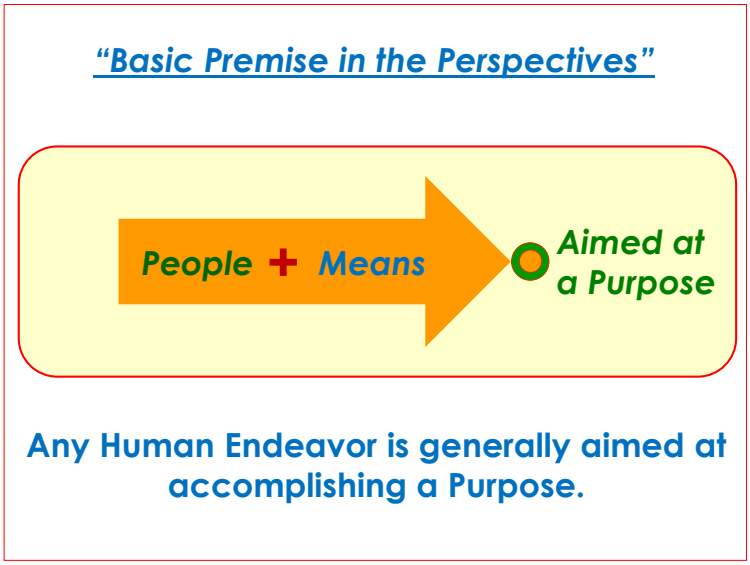


Fig 1 Basic Premise in the Perspectives

Since engineering is about addressing the concerns of sustenance (of a means that would aid), while accomplishing a purpose, the engineering design objective is to come up with such 'means'or scheme of things, that supposedly aid in accomplishing a purpose. As per systems methodology by J. Gharajedaghi [1], these means are the structure & process and knowledge is the underlying ingredient.

The various books of knowledge which supposedly describe the concerns a given 'means' will be subjected to, are portrayed as the possible solution space. We choose various schemes that aid 'Organizing Knowledge' as a case towards knowledge structure and discussed if it reflects the elements that we have described as part of this paper.

2 Systems Principles and Systems Methodology

According to J. Gharajedaghi [1], the following five principles act together as an interactive whole, and define the essential characteristics about systems. The five principles are:

- Openness
- Purposefulness
- Multidimensionality – partially included for discussion in this paper.
- Emergent Property – not included for discussion in this paper.
- Counterintuitiveness – not included for discussion in this paper.

Also as per J. Gharajedaghi [1], function-structure-process-context forms an inevitable whole towards a systems methodology. Accordingly at a smaller scope, function is equated with the purpose and carrying out a function with the aid of corresponding structure & process leads to accomplishing the purpose. This statement also means that in the absence of an associated process, structure alone will not be able to accomplish the purpose. This overall understanding is captured in table 1.

Concerns of Knowledge / Means for accomplishing a Purpose	Concerns of Knowledge that are *in favor* while accomplishing the Purpose	Concerns of Knowledge that are *not in favor* while accomplishing the Purpose
Means in the form of a *Structure*		
Means in the form of a *Process*		

Table 1. Means & Concerns

This paper is about our attempt to apply these principles and to put together our understanding about systems methodology, in order to arrive at a representation that can aid in engineering of knowledge structures. The systemic principles come into the context of our attempt, leading to the representation that we could arrive at. As an illustration, the various schemes that aid in 'organizing knowledge' with their respective structures are viewed through this generic representation.

3 Perspective from the Systems Principle of 'Purposefulness'

According to J. Gharajedaghi [1], one of the principles of systems is 'purposefulness' and he refers to human beings as purposeful systems. According to Russel Ackoff [2], a way to look at human behavior is to view them as systems of purposeful events. Hence a human endeavor ideally is aimed at accomplishing a purpose, whatever the purpose may be. Accordingly 'purpose' forms an important element in the representation towards engineering of knowledge structures.

A purposeful system is one that can produce not only the same outcome in different ways in the same environment but different outcomes in both the same and different environments. The 'scope of a purpose' can vary based upon the level at which one choose to operate. The discussion on what is a purpose and why is it a purpose is not part of the scope in this paper. In order to accomplish a purpose, people will leverage certain 'means' that supposedly adhere with systemic principles. The 'means' plus the people who would employ the 'means' to accomplish a purpose, is shown in figure 2.

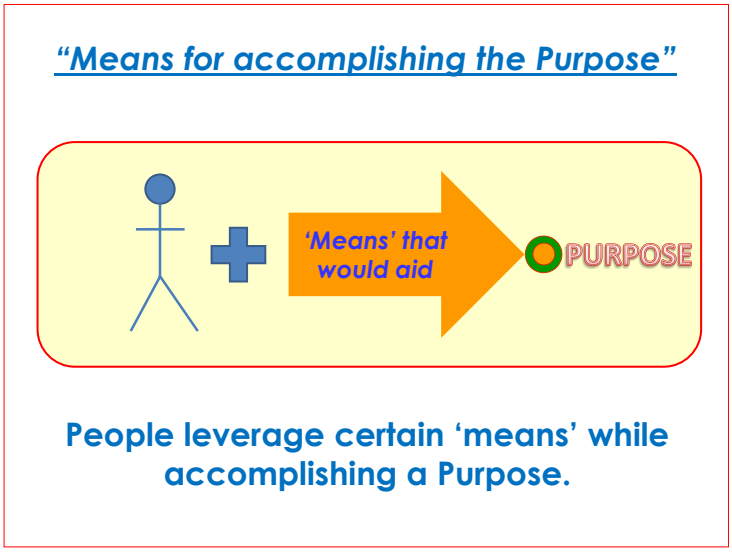


Fig 2 Means for accomplishing the Purpose

4 Perspective from the Systems Methodology

As per systems methodology, structure, function, and process with the context, define the whole or make the understanding of the whole possible. Structure defines components and their relationships; function defines the outcomes; process defines the sequence of activities; context defines the environment in which the system is situated. As per J. Gharajedaghi [1], iteration is the key for understanding the system and iteration on structure, function and process in a given context would establish the validity. Accordingly, the means amounts to the structure & process as depicted in figure 3.

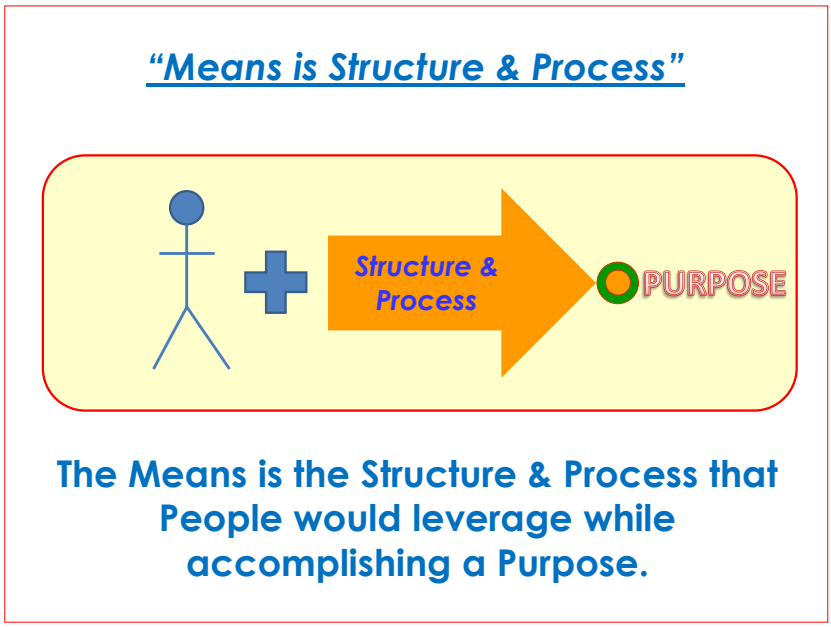


Fig 3 Means is Structure & Process

5 Perspective from the Systems Principle of 'Openness'

While attempting to accomplish a purpose, the forces or concerns that exist in a context or environment needs to be dealt with. These may be termed as 'influences' also. Accordingly, the associated knowledge that becomes relevant to the given purpose can be separated in to two: one is the knowledge that is 'in favor' of the purpose and the other is the knowledge that is 'not in favor' of the purpose. The same has been depicted in figure 4. The means that is employed to accomplish the purpose should be able to cope with these respective influences. For instance, a combined discipline of knowledge is a concern that 'means' aimed at 'organizing knowledge' should be able to cope with.

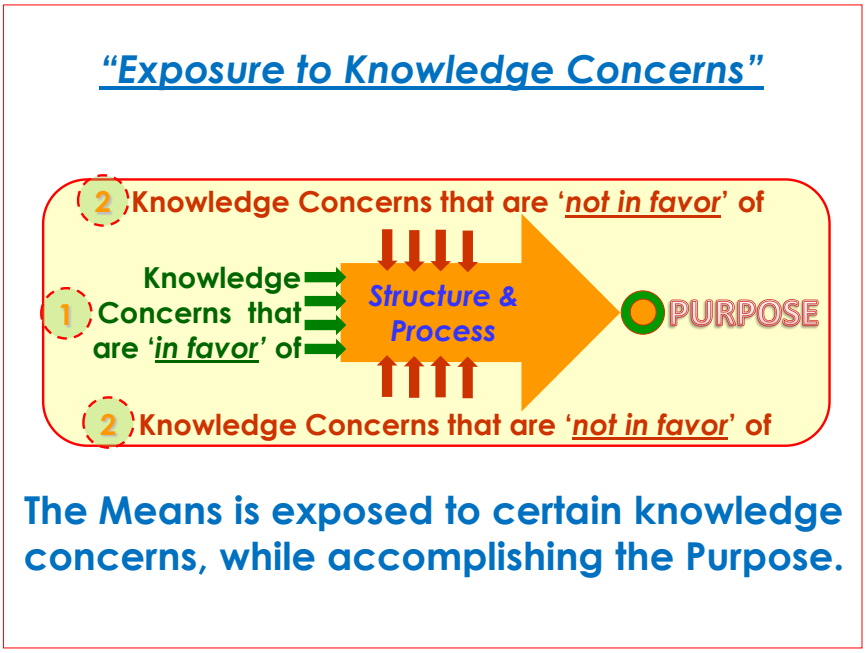


Fig 4 Exposure of Means to Knowledge Concerns

6 Perspective on Solution Space for the 'Means'

Openness also lends a handle towards access to various bodies of knowledge where one can find not only the knowledge of the problem but also knowledge of the solution too. The BOK (Book of Knowledge) consists of knowledge which refers to both of the concerns that are in favor of and also not in favor of, while accomplishing a purpose. Accordingly, the respective solutions also may be found within this knowledge base. The representation of it is depicted in figure 5.

Though we have separated the knowledge into two, relative to the purpose, the sources for identifying the knowledge are the same. They are the various bodies of knowledge resident in books and other media and also the body of knowledge that is present with people. These sources of knowledge consist of and refer to both those concerns (and influences) that are in favor and also not in favor, as depicted in figure 4. These sources of knowledge should not be confused with the 'classification of knowledge' which has been described as part of the illustration on 'organizing knowledge'.

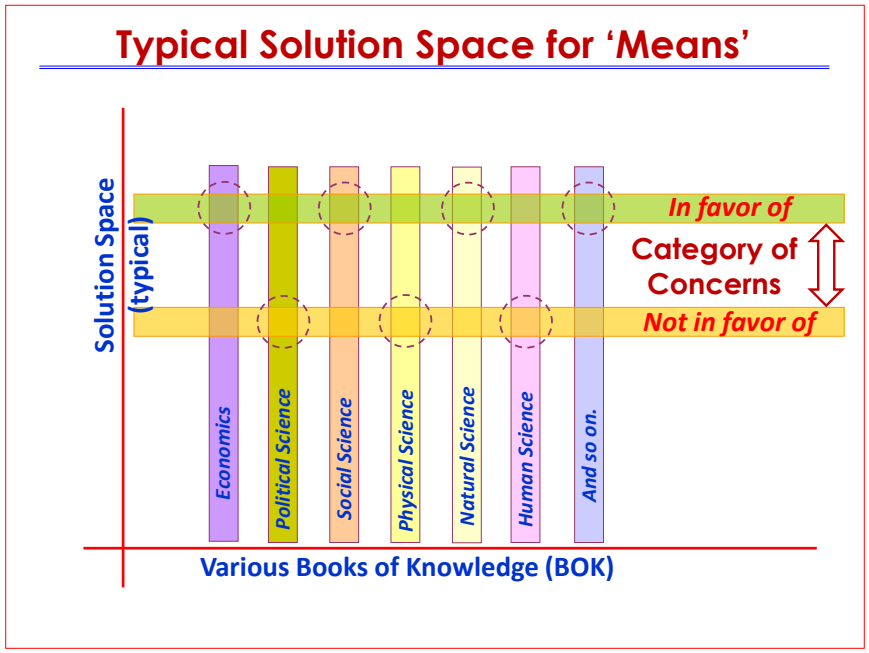


Fig 5 Typical Solution Space

7 Perspective from the Systems Principle of Multidimensionality

Multidimensionality is about the ability to see complementary relations in opposing tendencies. This principle maintains that the opposing tendencies not only coexist and interact, but also form a complementary relationship.

According to J. Gharajedaghi [1], "human beings form varying relations with each other, creating an interactive type of structure. Interactions between purposeful people in a group take many forms. People may cooperate on one kind of tendencies, compete over others and be in conflict over others and all of this at the same time. People learn and mature over time and are subject to change. The result is an interactive network of variable members with multiple relationships, recreating the network on a continuous basis. This is what is meant by plurality of structure".

In the context of knowledge structures, with knowledge being resident with people, accepting the plurality of structure is necessary to appreciate the principle of purposefulness and multidimensionality. Unlike traditional (physical) structures that endure, knowledge structures can continuously change and recreate themselves on a continuous basis.

Plurality of structure is an attribute by which 'knowledge structures' can recreate themselves continuously. As the purposes are realigned and the respective concerns change, the structures need to adjust towards renewed missions. Knowledge being non-physical, this manner of recreation is possible to achieve on an ongoing basis.

8 Further work on the Perspectives

Do we engineer the structure (and the process)? Or, do we engineer the knowledge? Perhaps we need to do both. What are the 'Degrees of Freedom'of the means employed? What are the limits of tolerance for them? What are the units or dimensions for knowledge structures? How do we take advantage of 'plurality of structure' in order to cope with change? How about the other principles namely counterintuitiveness and emergent property? What are the various realization levels [5] of discourse? All these are of interest and going forward we will study them. The proposed paper presentation can offer an opportunity for people to collaborate and carry out further study on these and related topics.

9 Context of 'Classification of Knowledge' and Description of various Schemes for 'Organizing Knowledge'

There are several schemes for organizing knowledge. Those who have come up with these schemes have thought about the various concerns and also they have thought about these concerns differently. What should be the ideal scheme? What are all the concerns that it should address? Why is it that a given scheme has become popular in spite of other schemes having better features?

We choose this illustration for the reason that it is to do with knowledge. The classification of knowledge has several schemes with corresponding structures aimed at 'organizing knowledge'.

There are various ways knowledge classification has been done by the research scholars in the world, Barbara H. Kwasnik [3] & A N. Raju [4]. These classification schemes are commonly used in the libraries in the world. There are various classification methodologies which are in place. Some of the popular ones are Dewey Decimal Classification (DDC), Universal Decimal Classification (UDC) and Colon Classification (CC). These are briefly described in the following paragraphs.

DDC has been devised by Melvil Dewey in the year 1876. It is the most widely used classification scheme available in about 135 countries in the world. It has been translated in to more than 30 languages in the world. More than 90% of the libraries in the world are using DDC for the classification of books, which includes Public, Academic, and Special libraries in the world. DDC has 7 standard subdivisions, representing areas, individual languages and literature, racial and ethnic groups, languages and persons.

The entire knowledge as per DDC has been classified between 000 to 999. These main classes are divided further into 10 main classes. Each class is subdivided into ten more and so on. Arabic numerals are used in the classification scheme to denote a given subject. The scheme also has the flexibility to assign new emerging areas of knowledge into the scheme. This means there are unassigned numbers throughout the scheme for the emerging subjects to be included into the scheme. 23rd version of the DDC was released in 2011. The DDC scheme has more orientation towards languages and literature.

The classification numbers are readily available to assign. The notations are arranged hierarchically, which will represent a base subject classification number. For example a book on *Indian Economics* is classified as *330.954* in DDC.

300	– Social sciences
330	– Economics
330.954	– 954 is for India taken from standard subdivisions Table 2 of DDC

UDC scheme is devised by Paul Otlet and Henri La Fontain at the end of 19th Century. It is an analytic synthetic classification scheme. This method of classification helps in indexing and retrieving information easily and faster. Though UDC has been devised based on the principles of DDC, it used more connecting symbols to represent a class number. The advantage with the UDC scheme is that it will enable us to go to minute level of detail, which is required for indexing and abstracting services. Approximately 3% of the libraries in the world uses UDC scheme for the classification purposes. It is mainly used for indexing articles in journals.

UDC is hierarchically expressive, which means the longer the number, specific the class. It also has a syntactical representation which means UDC codes are combined with the help of a COLON (:) to represent a two notational elements /subjects. It is the most flexible scheme of classification of knowledge. The scheme has more emphasis laid on social sciences and technological areas. The uniqueness of this scheme is, it has common auxiliary tables which it will represent with the help of various notations to represent places, people, races, medium etc. These auxiliary tables will facilitate to provide a specific classification number to a given area of knowledge. The scheme also has a provision to accommodate new and emerging subjects.

A book on *Indian Economics* is classified as **33 (540)** using UDC scheme. The classification number is arrived as below:

33	Economics broad subject
33.(540)	for India taken from Common Auxiliary Table of UDC

CC is popularly called Colon Classification, devised by SR Ranganathan, in the year 1933. It was the first faceted analytico-synthetic classification scheme. The name Colon comes from the punctuation mark COLON (:), which helps to separate the two facets of a classification number. Colon Classification number used 42 main classes, which includes letters, numbers and punctuation marks.

The entire knowledge structure in CC are classified alphabetically from A to Z, A for generalia to Z for law. In the lines of standard subdivisions in DDC and auxiliary tables in UDC, CC used PMEST to represent various facets. The CC scheme uses 5 facets which are called PMEST, which means Personality, Matter or property, Energy, Space and Time.

The CC is most user friendly scheme of knowledge classification scheme to assign exact class number for a given subject. PMEST is represented as:

Personality	(,) Comma
Matter	(;) Semi Colon
Energy	(:) Colon
Space	(.) Period/full stop
Time	(‘) Apostrophe

The classification scheme uses PMEST to complement and supplement various knowledge elements to arrive at a more clear class number to a title. Below illustration helps us to understand the construction of classification number using CC. For example a book on *Indian Economics* is classified as – X . 44 using CC.

X	Economics
.	Space for connect Geography India from the title
44	is for India from PMEST

10 Illustration in the context of ‘Organizing Knowledge’

One topic of illustration that we have identified for discussion in the proposed paper is ‘organizing knowledge‘. Now if we look at the scheme DDC from the point of accomplishing the purpose of ‘organizing knowledge’, the concerns or influences this particular scheme will be exposed to while accomplishing its purpose could be several. One such concern is creation of new knowledge. DDC scheme addresses this concern by way of having unassigned numbers which can be used by emerging knowledge.

The universe of knowledge is classified into broad 10 areas starting from 000 to 999, as per DDC. Any subject in the world finds its place in the classification scheme. Classification facilitates to store and retrieve information easily. Since the knowledge is classified into 10 main areas, the information we are seeking will help to relate to its broad subject area, that we can look for.

Each scheme is oriented towards a particular area. DDC is more oriented towards language and linguistics, literature and religion etc, where as UDC is more oriented towards social sciences and technology. The CC scheme helps to classify the title to the last level of detail. The CC scheme uses notations and punctuation marks to provide clear and distinct call number. This leads to lengthy *call* numbers.

In DDC, it will be difficult to assign a *call* number for a title which is dealing with more than one subject. This limitation is adequately addressed in UDC and CC while providing the use of colon and punctuation marks. UDC also uses decimals to denote a particular subject, and helps to arrive at complete details of a given title. Two subjects can be easily represented while separating with a colon or circular bracket.

DDC scheme has wider acceptance level, but if a book is dealing with more than one broad subject, the book can be classified under any one broad subject Depending on the priority you would like to assign. We cannot assign two broad subjects for a given title using DDC. The mechanism to connect two subjects is not available, unlike in UDC and CC.

Whereas UDC has overcome the challenge which DDC has, i.e we can classify a book dealing with more than one subject, by using a colon, to distinguish two broad subjects. The colon signifies representation of two subjects in a given title. For example, a book on *Political Jurisprudence* in India is classified as *34:32:934***.** In this 34 represents Law Jurisprudence, 32 for Political Science and 934 for India geography**.**

Colon classification scheme also helps to classify a book with more than two subjects by providing equal priority to two subjects, which is based on subject denoted classification scheme. An illustration of the same is shared here. For example, cultivation of Mangoes by applying financial viability is classified as J382:7 (X). The same can also be respresented as X382.7 (J) to lay emphasis on economics.

J	Agriculture
382	Mangoes
:	Energy
7	Cultivation
(X)	Economics

In the above example, two main subjects are brought together and represented equally.

The basic purpose of various classification schemes is to organise knowlege present in various forms in a more methodical manner, so that it helps in classifying in a logical way and arranged on the shelves for easy retrieval.

In todays world, knowldge is available in various forms like books, journals, periodicals, magazines etc. Knowledge creates more knowledge. It is challenging if this published knowledge is not arranged meaningfully. Various classifications schemes which are devised by Melvil Dewey, and others will help classify knowlege and arrange it on the shelves so that retrieving the knowlege is easier, without chaos.

While devising the classification schemes, the respective *designers* of the schemes have taken additional care to foresee the future demands and the new and emerging subjects that would come up. Accordingly they left the classification numbers to be accommodated into the existing classification scheme meaningfully and logically. These new emerging subjects take their logical position in the classification scheme, without disturbing the existing population of the scheme.

Basically all these schemes have been developed with the intention to arrange the knowledge so that it will be easy to retrieve as and when required. The various schemes that we have described are the 'Knowledge Structures' that are aimed at 'Organizing Knowledge'. If we look at these various schemes through the lens of a generic representation, the elements that ought to be present in each of the schemes are:

- Purpose of the scheme;
- Influences or Concerns the scheme is exposed to;
- Knowledge concerns in favor of and not in favor of the purpose;
- Plurality of structure in the scheme;
- And a few more which we have not yet discovered.

Accordingly the given scheme will either thrive or survive, depending upon the presence of these elements in the respective scheme. The proposed paper offers to discuss with active participation from the audience, so as to reinforce our understanding about engineering of knowledge structures.

11 Extending this work to Apply in the Context of other Knowledge Structures

There are several others contexts of human endeavors, be it managing knowledge, or enforcing a given knowledge or altogether a different connotation such as nurturing an organization culture or building a corporate brand and so on, in which the generic representation may be applied. Knowing 'knowledge as an entity' could be the key in all such endeavors [6].

Another example for a knowledge structure as per our understanding is 'culture'; it could be organizational culture or culture of a society or country. We very well find that technology is playing a key role in shaping up cultures. For instance use of mobile phones and other electronic communication devices have changed the manner in which people are connected and the way they communicate and relate with each other, whether it is younger generation or older generation.

Another example could be the affordability and earning capacity of individuals in shaping up cultures by way of giving rise to expensive living and credit bearing life styles. Phenomenon such as these can be studied and possibly influenced through the 'means' that are arrived at by taking into consideration the various influences that shape up these cultures.

Implications of the systems principles namely emergent property and counterintuitiveness have not been looked into, in the current scope of the paper. But these principles as well contribute in arriving at the representation that has been attempted. Technology can play a vital role in these endeavors by way of the enablement it can bring in shaping up the 'means'.

12 Summary

This paper is about presenting the views or perspectives from traditional engineering disciplines and principles of systems. The views that are proposed for presentation will help in undertaking engineering of knowledge structures. The basic premise in this paper is that people leverage a certain means in order to accomplish a purpose. The means is stated to be the structure & process, both of which will get exposed to concerns that are in favor of and also not in favor of accomplishing the purpose. The engineering design objective is to build enough sustenance into both the structure & process so as to cope with these concerns. The solution space may be found in various books of knowledge where the concerns also are described.

The chosen illustration can be discussed from all possible angles so that our understanding about engineering of knowledge structures is strengthened. It is our sincere hope that these perspectives will be of help to the audience in their efforts to engineer knowledge or knowledge structures.

13 Acknowledgements

The authors are grateful to Jamshid Gharajedaghi for the knowledge and wisdom that he has presented about systems. The views presented in this paper are not possible without his inspirational writings.

Prof. Kesav V Nori, who has shaped our thoughts while leading business systems research, is quite instrumental in our attempt to bring out this particular paper.

References

1. Jamshid Gharajedaghi, Systems Thinking, Managing Chaos and Complexity, Second Edition, Butterworth-Heinemann, 2006, Elsevier Inc.
2. Russel L Ackoff, Fred E Emery, On Purposeful Systems, Transaction Publishers, New Jersey, 2008
3. Barbara H. Kwasnik, The role of Classification in Knowledge Representation and Discovery, School of Information Studies, Centre for Science and Technology, Syracuse University, Syracuse, NY, 1999.
4. AN Raju, Library Classification Theory, Himalaya Publications
5. Doji Samson Lokku, Kesav V Nori, Morphogentic Constraint Satisfaction Based Approach for Organizational Engineering, ACM SAC, Dijon, France, 2006
6. Doji Samson Lokku, Importance of knowing ‘knowledge as an entity‘ in order to contribute towards Service Science Education, People Education Congress, HBCSE, TIFR, Mumbai, 2009

Engineering of Dynamic Knowledge in Exact Sciences: First Results of Application of the Event Bush Method in Physics

Cyril A. Pshenichny[1], Roberto Carniel[2] and Paolo Diviacco[3]

[1] Geognosis Project, Intellectual Systems Laboratory, National Research University of Information Technologies, Mechanics and Optics, Kronverksky Prospect, 49, St. Petersburg 197101, Russia

cpshenichny@yandex.ru

[2] Laboratorio di misure e trattamento dei segnali, DICA, Università di Udine, Via delle Scienze, 206, 33100 Udine, Friuli, Italia

roberto.carniel@uniud.it

[3] Istituto Nazionale di Oceanografia e di Geofisica Sperimentale, Borgo Grotta Gigante 42/c, 34010 Trieste, Italia

pdiviacco@ogs.trieste.it

Abstract. The tools of knowledge engineering are commonly applied in poorly formalized information domains. However, there is a number of reasons to use them also in exact sciences along with mathematics and mathematical logic. In physics, the knowledge is largely dynamic, i.e. describes rather processes and changing objects than objects with fixed properties and states. Therefore, even ontologies in physics tend to include a "dynamic" component. Nevertheless, application of specific methods of dynamic knowledge engineering looks very promising and beneficial for physics. The paper reports first results of application of the event bush method in theoretic and applied (geophysical) contexts.

Keywords. Physics, knowledge engineering, dynamic knowledge, event bush method, ideal gas, site effect

1 Introduction

The sense of application of knowledge engineering tools in exact sciences may look arguable to many researchers. Indeed, the virtue of knowledge engineering is to shape up and structurize the semi-intuitive fields of knowledge, extract axioms and suggest inference where these are unknown. Physics and chemistry, being still far from as

formal as, say, geometry, are nevertheless well expressed in terms of mathematics and seem to require only mathematical logic for further (and perhaps complete) formalization.

Still, there are a few reasons to think differently.

1. To make physical or chemical issues computer-understandable, one needs to convey meaning to information systems, while the latter requires knowledge-engineering technologies. As was shown by Borst et al. [3], computer may readily recognize the mathematical operation of multiplication in an expression $F=ma$ but it is a highly non-trivial task to make it understand this expression as a part of Newtonian mechanics.
2. It often becomes necessary or desired to explain the matters of physics or chemistry to a wider community, e.g. to make geophysical models clear to other geoscientists [3; 11]. A method enabling one to share knowledge with a wider community (perhaps even without sharing it with a computer) is a pre-requisite for creation of computer-based collaboration environments.
3. The accuracy of existing and new physical models sometimes needs to be checked not only for mathematical errors but also to analyze whether the variables have been used adequately to reflect personal or collective vision of the phenomenon. This may strongly help to create new models and reconcile standpoints of physicists or chemists. Such reconciliation may appear especially helpful for physics. In physics, people often think first in terms of things and events (processes) and then convert their vision into variables and mathematical operations, and this passage is often determined by intuition or by existing bias in a given scientific school. Many authors, as for example Polson and Curtis [10] or Bond et al. [2] highlighted the role of previous experiences and preconceived notions that stem from their personal backgrounds. Diviacco [6] analyzed the relationship of creative (abductive) reasoning and social positioning of researchers and scientific institutes. Baddeley et al. [1] reported on the phenomenon of opinion shaping and herding, while suggesting paths to interrogate experts through elicitation.
4. Computer-aided engineering may benefit from “parsing” the physical models used in it [15].
5. Perhaps, knowledge engineering would offer new formalisms, which, along with existing mathematical and logical approaches, would also suit these fields well and be useful to them. Thus, one may expect that the method of event bush [8; 10] evolves into such formalism with time.

These considerations have led to a number of applications of knowledge engineering methods in physical and chemical domains.

Like in other domains, two tendencies can be recognized here, static and dynamic knowledge modeling. By static we mean the representation of information domain as a no-change environment sensu Pshenichny and Kanzheleva [13]. Static knowledge engineering proceeds in exact sciences, like in other fields, mainly by means of ontology design. The same time, the character of knowledge in exact sciences, especially in the physics which will be the focus henceforth in this paper, is such that even the properties of objects (e.g., mass, momentum, charge or energy of a body, orientation and intensity of a field) are considered mainly in relation to possible response to some

external impact, i.e., dynamic view is implicitly wrapped even in formally "static" descriptions. This is why it is not surprising at all that the ontologies used in this domain often tend to incorporate a "dynamic component". Though, the ontology proper offers not too much to capture processes, the solution found by Borst et al. [3] is to create an ontology of processes as chains of flows, define flows as changes of particular kind of stuff (mass, energy, location, charge and so forth), link this stuff to components of physical system, associate these components with variables and decompose them into sub-components, which, in turn, would correspond to various values of variables. Proceeding from one value of variable to another under specified conditions (the latter represent physical function defined by a model or law) is visualized by a bond graph [4]. Though only variables and variable states are in the nodes of the bond graph, if the latter is interpreted in terms of subcomponents between which some stuff changes, it should be regarded as an event-based tool. As such, the bond graph is the counterpart to the essentially static PhySys ontology.

An approach demonstrated by Yoshioka et al. [15] is rather similar to that of Borst et al. [3]. In the physical concept ontology they develop in support of computer-aided engineering they suggest the following "conceptual categories". Two of them describe the "static structure of knowledge" (entity, representing an "atomic physical object", and relation, meaning the "relationship among entities to denote a static structure"). Other three relate to "dynamic" part of physical knowledge: attribute ("a concept attached to an entity that takes a value to indicate the state of the entity"), physical phenomenon that "designates physical laws or rules that govern behaviors" and physical law "representing a simple relationship among attributes". Importantly, the passage from static to dynamic knowledge in both studies, in fact, means passage from names (of subcomponents in one case or attributes in the other) to variables, from the qualitative to quantitative vision. Only variables and their correctly built combinations (i.e., laws and models represented by formulae) are believed to be able to represent the behavior of what is called physical system by the quoted authors.

This is a rather common point in physical modeling resulting in the fact that when the event-based methods are applied in physics (bond graphs, equation graphs, flowcharts sensu stricto or sensu lato), quite often only variables, equations or, at best, questions occur in their nodes and meaning of arcs may be not defined at all. Perhaps from the point of view of physicists such semantics is perfectly organized as it gets rid of virtually anything qualitative. Still, if not to consider quantitative thinking as some upper stage of evolution of brain and take it as a part of existing scientific and cultural context, the semantics of these graphs and corresponding domains of physical knowledge looks rather groundless and needs substantiation in words, though, managed as strictly as variables.

In physics do exist visions that already imply their description by an event-based knowledge engineering method – e.g., branching process suggesting application of event trees (directed invariant change environment), various cyclic processes calling for causal loops (non-directed change environment), and so forth, up to the most complicated, alternative directed change environment.

However, these methods, as was argued by Pshenichny and Mouromtsev [13], contrary to the object-based ("static") ones, need better rethink and formalization of grammar of event description in nodes.

One method of dynamic knowledge engineering, the event bush, has got the most evolved verbal grammar and, besides, addresses the most complicated and all-inclusive modeling environment, that of alternative directed changes [13]. In our paper we will examine an opportunity of formalization of record of events, processes and scenarios in physics by means of the event bush method.

To do this, first we consider a trivial task from theoretical physics, and then review a geophysical application of the event bush.

2 Theoretical Task

The theoretical task we considered for application of the event bush represents one of the fundamental physical laws known as Mendeleev-Clapeyron Law, or ideal gas equation. Being used in thermodynamics, statistical mechanics and kinetic theory of gases, it links pressure, volume, temperature and amount of gas. Its most frequently introduced form is

$$pV=nRT \tag{1}$$

where p is the pressure of the gas, V is the volume of the gas, n is the amount of substance of gas (also known as number of moles), T is the temperature of the gas and R is the ideal, or universal, gas constant, equal to the product of Boltzmann's constant and Avogadro's constant.

To put the knowledge from equation (1) into an event-based framework one needs to understand how the reasoning in terms of parameters corresponds to reasoning in terms of events. In particular, this means to relate variables of the equation to some propositions, their subjects and predicates.

For this, we shall conceptualize this equation as a possible scenario in some environment of directed alternative changes, in which we first of all have to define, as implied by the event bush method, some "key players" being primary, non-unique inputs, which would determine any further course of events, and external "actors" that may put some constraints on the behavior of "key players" [14]. Terms defining "key players" become subjects of primary internal, and those defining external actors, the subjects of primary external events in the event bush.

In our case, it is reasonable to look for those "key players" and external "actors" whose properties can be expressed as variables in the considered equation. R should be excluded from consideration because it is a constant. What remains is pressure p, volume V, number of molecules (moles) n and temperature T. If to take molecules of gas as "key players", then their amount is obviously their property, temperature is a characteristic of their chaotic motion, and pressure is also their characteristic, of the overall impact they exert on something. Likely, exactly this "something" is an external "actor", whose appearance gives sense to the pressure. And, obviously as well, its

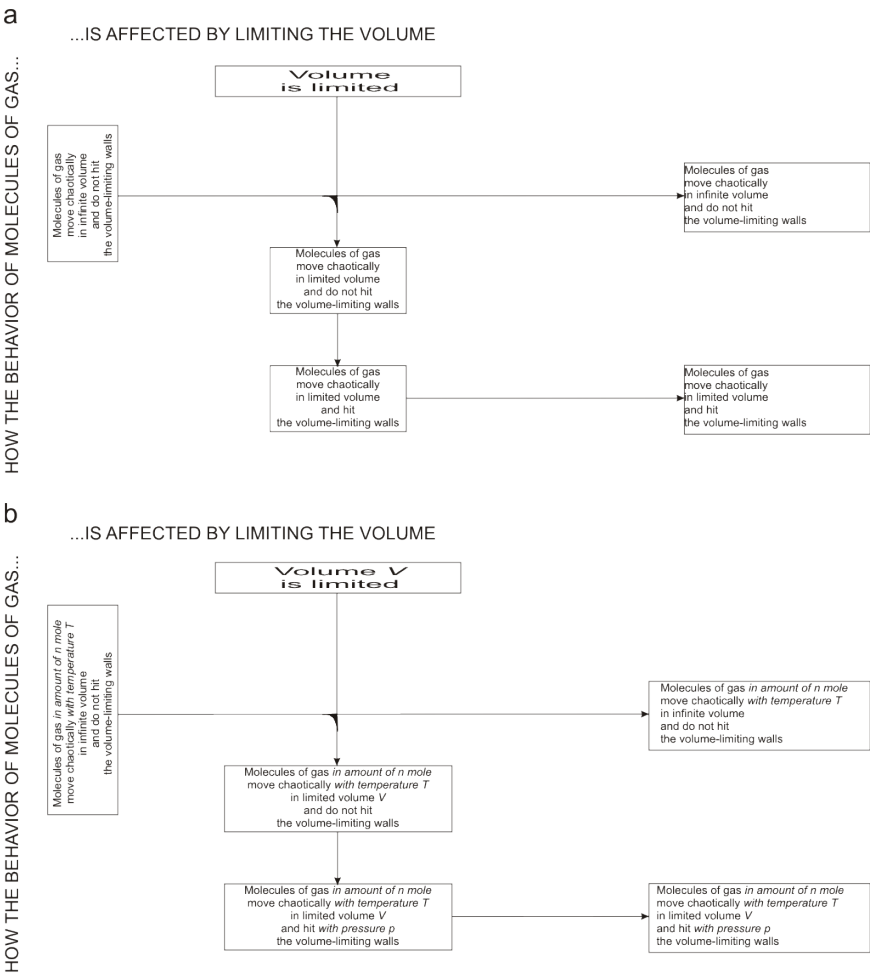
property must be “to have volume” or, to put it more correct, to have finite volume, *V*. However, as all the variables have been included into consideration, it will be good if neither we have more “key players”, nor other external “actors” are needed, nor any other properties (predicates) are needed for this only external “actor” we are trying to define. In fact, “to have volume” and “to be finite” is all we need from it. Then, whatever it be, we can call it just “volume” and attribute the predicate “to be finite” to it. However, “to be finite” may be axiomatically taken equivalent to “to have walls”. This predicate can be used to constrain the definition of “key players” before and after the action of the external “actor”.

Then, the structure of the event bush appears as follows: (ia) “Molecules of gas move chaotically in infinite volume and do not hit volume-limiting walls”; (ib) “Volume is limited”. The resulting bush is shown in Fig. 1, a. Importantly, when external “actor” comes to play, i.e., the volume in which molecules of gas reside *becomes* limited (no matter in what way), we cannot say that molecules of gas *immediately* exert pressure on the volume’s walls. No, this takes some, however short, time, though happens inevitably and without any additional influence. Semantically this is expressed in the change (ii) “Molecules of gas move chaotically in finite volume and do not hit volume-limiting walls” FLOW (ii) “Molecules of gas move chaotically in finite volume and hit volume-limiting walls”.

When the bush is ready, it looks more or less straightforward to relate variables to its events (see Fig. 1, b). Each event in event bush is a statement consisting of one subject and whatever number of predicates. If a subject or predicate are quantifiable (e.g, a subject “molecules of gas” may imply number of molecules, predicate “to move chaotically” implies temperature that characterizes this movement), this quantitative parameter may be readily added to the formulation of the event right after the corresponding subject or predicate – “molecules of gas *in amount of n moles”* or “move chaotically *with temperature T*”. We put the parameter in italics for clarity. Thus a qualitative parameter both formally and stylistically “stems from” the qualitative formulation of event. In some cases, moreover, event’s subject or predicate may be a parameter to itself, e.g., “volume”. Once attributed to a subject or predicate, a parameter is traced throughout the bush in all events in which this subject is present or this predicate is asserted. If the predicate is negated, no parameter is associated with it.

When parameters are associated with events, each connective of the bush is being attributed a computational sense (Fig. 1, c) serving as a step in construction of some formula (one of the formulae constructed by the bush represents the considered law). Either a connective adds one mathematical operation over the variables present in the nodes (events) it connects, or it presents an assumption relevant for the formula. Herewith, it becomes important again that semantic peculiarity in change between two (ii) events fixed by the flux connective, which was quoted one paragraph above. Now it is laden with an assumption that puts in explicit form the fact that we ignore the time elapsing after the volume becomes closed and before the pressure appears in it.

Following the two flows of the event bush, we arrive either to a “nothing happened” case or to the scenario that is described by the Mendeleev-Clapeyron law and see the relevant physical formula is being constructed.



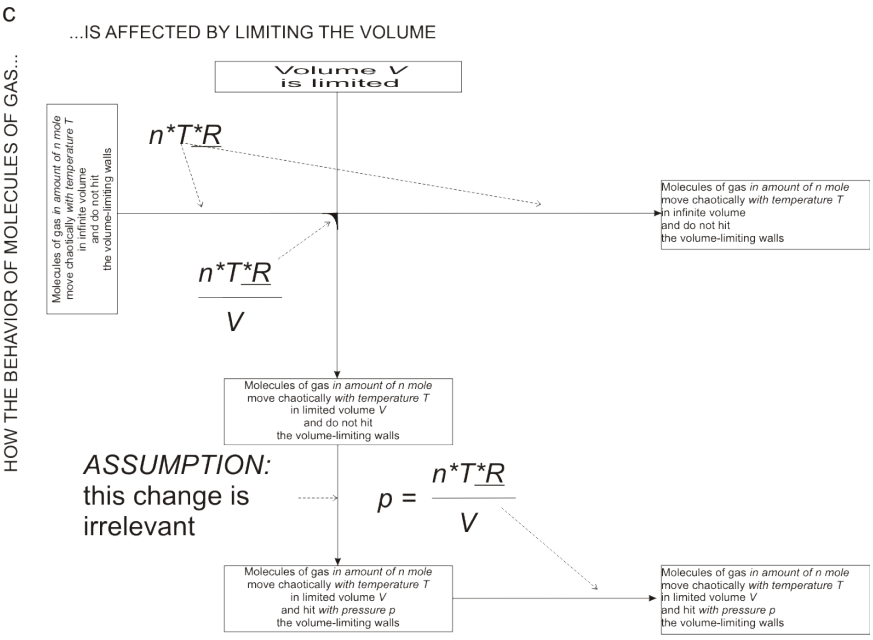


Fig. 1. An event bush describing the environment, whose one scenario is modeled by the Mendeleev-Clapeyron law: (a) environment proper, (b) environment with variables attributed to events, (c) environment with variables, assumptions and mathematical operations. See how the formula of the law is built step by step in event bush.

3 Geophysical Application

The same approach of construction of/"parsing" the physical formulae, or models, was used to explicate one seismological model of the so-called site effect.

Many earthquakes have indicated that the presence of deposits of soft soil over an underlying harder rock can increase dramatically and/or concentrate locally damages and life losses. Soft soils amplify shear waves and, thus, amplify ground shaking. This amplification of motion over soft sediments is called site effect and takes place mainly due to the trapping of seismic waves associated to the impedance contrast between the more superficial sediments and underlying bedrock.

The horizontal-to-vertical (H/V) spectral ratio approach, i.e. the study of the ratio between the amplitude spectra of the horizontal and the vertical component of seismic noise, was first introduced by Nogoshi and Igarashi [9] but became widely known only with the work of Nakamura [7]. The same Nakamura then clarified his method in a more recent paper [8].

Horizontal and vertical spectra on the surface ground of the sedimentary basin (H_f, V_f) can then be written as follows:

$$H_f=A_hH_b+H_s;\ V_f=A_VV_b+V_s \tag{2}$$

where A_h and A_v are the spectral amplification factors of the horizontal and vertical motion of vertically incident body waves, correspondingly, while H_b and V_b are the spectra of horizontal and vertical motion as they would be acquired in the bedrock under the basin. H_s and V_s are the spectra of horizontal and vertical components respectively of surface waves traveling along the earth surface. Thus two types of seismic signals are involved in producing the final site effect, called by us for simplicity "interior spectrum" (i.e. a seismic signal with given – horizontal or vertical – spectrum, originating in the earth's interior) and "exterior spectrum" (i.e. a seismic signal with given – horizontal or vertical – spectrum, coming to the surface from anywhere on or above it).

So the entire model of site effect on seismic spectra as suggested by Nogoshi and Igarashi [9] can be represented as $H_f/V_f=(H_b/V_b)*((A_h+H_s/H_b)/(A_v+V_s/V_b))$ – being is a purely mathematical derivate from two equations, $H_f=A_hH_b+H_s$ and $V_f=A_VV_b+V_s$, which are actually the model. Exactly these two equations will be addressed by the event bush.

For more detail, the reader is referred to the paper by Carniel et al. [5].

To parse equation (2), an event bush was composed that describes how the propagating seismic spectra are transformed by heterogeneous geologic substrate up to the earth surface (Fig. 2) based on the qualitative understanding of this phenomenon.

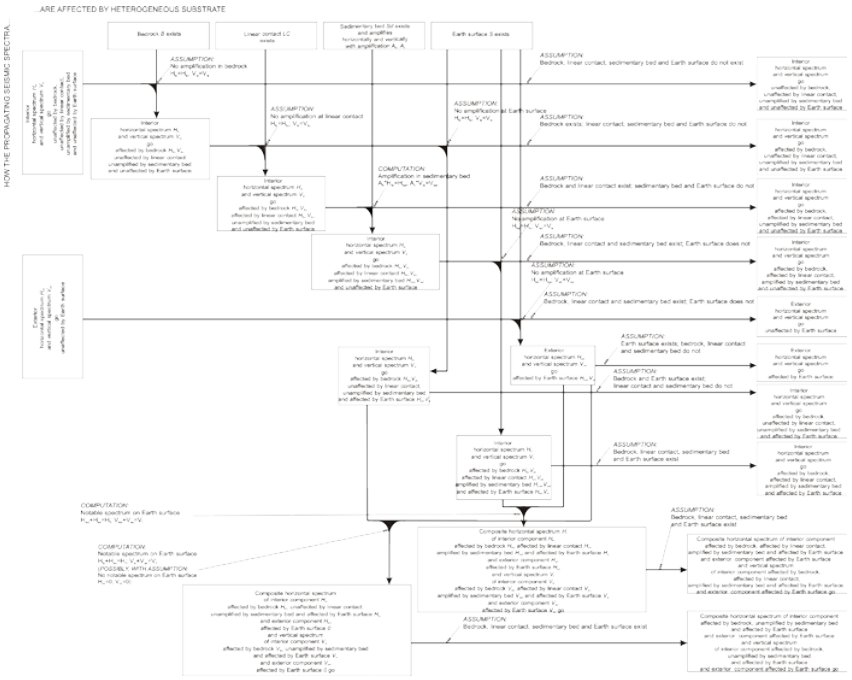


Fig. 2. Event bush describing how the propagating seismic spectra are affected by heterogeneous substrate

As is evident from the title of the bush, its main "players" should be the interior and exterior couples of horizontal and vertical seismic spectra (initially placed on the left, i.e., being classified as primary internal events) and geologic substrate (bedrock and sedimentary bed) plus the Earth surface, which are put on top of the bush as primary external ones.

As could be concluded from the Nogoshi – Igarashi – Nakamura model, the interior couple is considered to come to the bedrock but is not generated in it, because the opposite would cause other, *side effects* that could appear influential on propagation of spectra and had to be accounted for. To avoid this unnecessary complication, as well as for the sake of semantic clarity, we consider interior couple of spectra per se as a primary internal event. Its full formulation is "Interior horizontal spectrum and vertical spectrum go unaffected by bedrock, unamplified by sedimentary bed and unaffected by Earth surface". By "affected" we mean any kind of influence, i.e. a very general concept of (possibly nonlinear) filtering. The model says that the sedimentary bed *amplifies* spectra (and this is its influence on the latter, mathematically representable with a linear filter transfer function); however, we may not exclude that other two primary external "players", bedrock and surface, also may *affect* (or not) somehow. So we decide that the predicate "to affect" is attributable to any subject including "Bedrock" and "Earth surface", while the predicate "to amplify" is attributed to "Sedimentary bed" from the very beginning.

In principle, interior spectra may never meet the bedrock *meant in the model* (say, traveling yet deeper than, or far away from, this bedrock body all along). Exactly this is what the same-formulated tertiary event caused by this primary internal one is introduced for. Hence, the encounter of *that* bedrock by the spectra is another, secondary event "Interior horizontal spectrum and vertical spectrum go affected by bedrock, unamplified by sedimentary bed and unaffected by Earth surface". Naturally, this secondary event is caused by the said primary internal and a primary external one, "Bedrock exists".

From this point, other possible events involving interior spectra develop. These spectra can either travel in the considered area within the bedrock and not leave it (this is expressed by a tertiary statement formulated the same way as the above secondary one), or come to sedimentary bed, or come to the surface. Two latter options are secondary events that result from a combination of the secondary event "Interior horizontal spectrum and vertical spectrum go affected by bedrock, unamplified by sedimentary bed and unaffected by Earth surface" with a primary external, "Sedimentary bed exists and amplifies", in one case, and "Earth surface exists", in the other. Exactly these two secondary events represent the contrasting cases analyzed by Nakamura. Each of them can be clearly documented, i.e., represent some end results of propagation of the spectra, and this is reflected in the bush by corresponding tertiary events.

Now, another independent "actor" comes to play. This is a purely superficial, or atmospheric, or anthropogenic, event that may also generate vertical and horizontal seismic spectra, henceforth denoted as "exterior". This is another primary internal event of the bush. In principle, it may pass unrelated to the interior spectra, just "meaning itself" and resulting in a tertiary event. Nevertheless, if a portion of it

comes to the Earth surface (i.e., a combination of this primary internal event with the primary external one "Earth surface exists" takes place), this results in a couple of superficial spectra spreading close to the ground and being somehow affected by the latter. This is expressed by the secondary event, "Exterior horizontal spectrum and vertical spectrum go affected by Earth surface", and the end result of this, by the same-named tertiary event.

If two different couples of spectra, one coming from the interior, the other purely exterior, meet at the Earth surface, this naturally leads to two interior-exterior "couples from couples", one interior-exterior horizontal and one interior-exterior vertical. These "couples from couples" will be denoted composite spectra. Their Interior components bear the history of previous transformations ("affected by bedrock, amplified or not by sedimentary bed and affected by Earth surface"), while exterior ones may only be affected by Earth surface. In the event bush, this is expressed as confluence of the event two events – one of these in both cases is in one case, "Exterior horizontal spectrum and vertical spectrum go affected by Earth surface", and the other, depending on what kind of interior spectra are involved (or, in other words, where the seismometer is located), either "Interior horizontal spectrum and vertical spectrum go affected by bedrock, unamplified by sedimentary bed and affected by Earth surface", or "Interior horizontal spectrum and vertical spectrum go affected by bedrock, amplified by sedimentary bed and affected by Earth surface". Both events describing the composite spectra lead to tertiary events, which document the two scenarios captured by the Nogoshi – Igarashi – Nakamura model.

Some important notes can be made on the overall structure of the event bush.

Its primary internal events include two different types of spectra couples, interior and exterior (see above). In the model examined, no other seismic signals are considered.

Primary external events include the members of the geological sequence (bedrock and sedimentary bed) and the earth surface. Again, in the considered case this is surely the full set of opportunities.

Secondary events (ii) fall into four classes:

1. those formed by combination of ia and ib;
2. those formed by combination of the 1st class type (ii) events and one event ib (namely, "Earth surface exists");
3. those formed by confluence of the 2nd class type (ii) events and one of the 1st class type (ii) events (namely, "Exterior horizontal spectrum and vertical spectrum go affected by Earth surface");
4. those formed by simple cause-effect relation from the 3rd class type (ii) events.

Tertiary events, as supposed by the rules of event bush composition, were generated by primary internal ones and secondary events except those of the 3rd class. Their formulation repeats that of the events they originate from.

We do not ascertain that this bush is the only possible one describing the transformation of seismic signals in heterogeneous geological environment. It would be interesting to try to build other bushes in a different semantics and look at their interrelation. Moreover, a "vice versa" bush can be created describing the way the geolog-

ic bodies (bedrock, sedimentary bed and, finally, the earth surface) are being affected by seismic spectra and, in turn, transform the latter.

Like in the case of Mendeleev-Clapeyron equation, the succession of steps for building the bush that fits physical model is the same: building the bush proper, attributing variables to the events, attributing mathematical operations or assumptions to the event bush connectives. Aiming to "parse" and clarify the Nogoshi-Igarashi-Nakamura model, we found several implicit assumptions that were meant but not mentioned by the scientists relating the amplification at linear contact and Earth surface. Some of the assumptions like "Bedrock and linear contact exist; sedimentary bed and Earth surface do not" are obviously non-realistic (i.e., there is no contact without sedimentary bed) but formally should be mentioned in general framework. Interestingly, those assumptions linked with flow connectives, which are not associated with influx, seem to have a kind of "a posteriori" meaning fixing the pre-requisites that *have been* used to infer an event. Also, we found it necessary to introduce a new and more elaborated system of designations of seismic spectra to avoid confusion present in the authors' formulae (e.g., it is not clear that the authors' H_s and V_s are meant to include the history of passing the bedrock, which is designated by other members of the formulae, H_b and V_b, and also include the "surface history" of some of the spectra, while this cannot be concluded from the formulae). Also, the event bush clearly discerns the cases of "qualitative zero" (e.g., "no amplification at linear contact" and, hence, no corresponding variable introduced in the bush) and "quantitative zero" (seismic spectrum on the surface may be present or not, so the corresponding variable is present but may be equal to zero). However, whether this can be considered a method and, if not, what should be done to make it such, will be discussed below.

4 Discussion

The two physical tasks considered above show that the proposed approach may work under some conditions, but this is definitely not enough to postulate that it is a method ready for extensive application.

The following issues, to our mind, deserve particular attention: whether this good, bad or irrelevant for physical modeling if not all subjects/predicates are attributed variables, whether different variables can be attributed to similar subject and predicates (i.e., one event bush gives birth to seemingly incompatible physical models), how to formally define those flows in event bush which really lead to construction of a formula (see Fig. 1c). A very important dimension of research is further elaboration of event description grammar that would hopefully allow one to avoid nonrealistic assumptions, which have been required by formal reasons. Perhaps, existing formalization is still not strong enough to easily formulate the rules of assignment of variables to events and operations/assumptions to connectives.

More such tentative studies should be carried out by various research groups to examine the conditions under which this approach works well. It should be noted here-

with that the reported event bush failed to incorporate other models of the site effect phenomenon suggested in the literature. Obviously, there remain unresolved problems in this approach and not all of them are even well understood.

However, one of such problems is possible multiple qualitative interpretations of similar formula/model, i.e., not only one event bush may "host" several models but the opposite may also be possible.

Another point that seems noteworthy is that different mathematical expressions are often known for a similar law. Probably these should be considered as different laws with independent qualitative interpretations or there a "qualitatively preferred" form of each formula – by analogy to that among many mathematical expressions of the same law there are those which make physical sense and those that represent merely mathematical reformulation.

Importantly, this approach should work not only to explicate the existing knowledge but also to build new models and suggest physical laws.

Further research is needed to clarify these points and introduce dynamic knowledge engineering as a method of physical research. Nevertheless, when developed, this method could work well not only in physics but also in other exact sciences where reasoning proceeds partly or entirely in terms of variables, e.g., economics, or technical design. Finally, when the semantics of event description is developed equally well as semantics of objects and relations in object-based methods of knowledge engineering, not only the event bush but other relevant approaches will become applicable in "parsing"/creation of physical formulae that would better capture particular patterns of reasoning – e.g., event tree or its ramifications, to address branching processes, causal loops, to mimic cyclic ones, and so forth.

5 Conclusions

1. Only the event-based methods may allow "parsing" of existing physical laws and models and creation of new ones based on commonly-shared qualitative reasoning.

2. To develop a method of "parsing"/creation of physical models and laws, there should be a semantics of events as strict and formalized as semantics of objects and relations in objects-based methods of knowledge engineering.

3. So far, the method of event bush seems to be the only candidate that fits the above requirements and can at least efficiently "parse" physical models and laws in some cases.

4. Further research is needed to find out its limits of applicability and formulate the rules of its use in physics.

References

1. Baddeley, M.C., Curtis, A., and Wood, R.: An introduction to prior information derived from probabilistic judgments: elicitation of knowledge, cognitive bias and herding, in: Geological Prior Information, Curtis, A. and Wood, R. (Eds), Geological Society, London, Special Publications, 239, 15-27 (2004).

2. Bond, C.E., Shipton, Z.K., Gibbs, A.D., and Jones S.: Structural models: optimizing risk analysis by understanding conceptual uncertainty, First Break, 26, 65-71 (2008).

3. Borst, P., Akkermans, H., and Top, J.: Engineering ontologies. Int . J . Human – Computer Studies, 46, 365 – 406 (1997).

4. Broenink, J.F.: Introduction to Physical Systems Modelling with Bond Graphs, http://www.ce.utwente.nl/bnk/papers/BondGraphsV2.pdf (1999)

5. Carniel, R., Pshenichny, C., Khrabrykh, Z., Shterkhun, V., Pascolo, P.: Modeling Models: Understanding of Structure of Geophysical Knowledge by Means of the Event Bush Method, in: IAMG Proceedings, Mathematical geosciences at the crossroads of theory and practice, Marschallinger, R., and Zobl, F. (Eds.), Salzburg, September 2011, 1336-1350 (2011).

6. Diviacco, P.: Addressing Conflicting Cognitive Models in Collaborative E-Research: A Case Study in Exploration Geophysics, in: Collaborative and Distributed E-Research: Innovations in Technologies, Strategies and Applications, IGI Global press, DOI: 10.4018/978-1-4666-0125-3.ch012 (2012).

7. Nakamura, Y.: A Method for Dynamic Characteristic Estimation of SubSurface using Microtremor on the Ground Surface: Q Rep Railway Tech Res Inst, 30, 25–33 (1989).

8. Nakamura, Y.: Clear identification of fundamental idea of Nakamura's technique and its application, in: 12th World Conference of Earthquake Engineering Proceedings, Auckland, New Zealand (2000).

9. Nogoshi, M., and Igarashi, T.: On the amplitude characteristics of microtremor (Part 2). J. Seismol. Soc. Jpn., 24, 26–40 (1971).

10. Polson, D., and Curtis, A. Dynamics of uncertainty in geological interpretation. Journal of the Geological Society, London, 167, 5–10 (2010).

11. Pshenichny, C.A., and Diviacco, P.: Descending to Misunderstanding in Collaborative Geoscientific Research Projects, in: IAMG Proceedings, Mathematical geosciences at the crossroads of theory and practice, Marschallinger, R., and Zobl, F. (Eds.), Salzburg, September 2011, 1365-1377 (2011).

12. Pshenichny, C.A., and Kanzheleva, O.M.: Theoretical foundations of the event bush method. In Societal Challenges and Geoinformatics, GSA Special Paper 482, Sinha, K, Gundersen, L., Jackson, J., and Arctur, D. (Eds.), 139-165 (2011).

13. Pshenichny, C.A., and Mouromtsev, D.I.: Representation of the Event Bush Approach in Terms of Directed Hypergraphs, in: ICCS Proceedings, International Conference on Conceptual Structures. Pfeifer, D., and Ignatov, D., (Eds). Mumbai, January 2013, in press (2013).

14. Pshenichny, C.A., Nikolenko, S.I., Carniel, R., Vaganov, P.A., Khrabrykh, Z.V., Moukhachov, V.P., Akimova-Shterkhun, V.L., and Rezyapkin, A.A.: The Event Bush as a Semantic-based Numerical Approach to Natural Hazard Assessment (Exemplified by Volcanology). Comp Geosc., 35, 1017-1034 (2009).

15. Yoshioka, M., Umedab, Y., Takedac, H., Shimomurad, Y., Nomaguchie, Y., and Tomiyama, T.: Physical concept ontology for the knowledge intensive engineering framework. Advanced Engineering Informatics, 18, 95–113 (2004).

Adjustment of the Event Bush Method to Chemical and Related Technological Tasks

Cyril A. Pshenichny

Geognosis Project, Intellectual Systems Laboratory, National Research University of Information Technologies, Mechanics and Optics, Kronverksky Prospect, 49, St. Petersburg 197101, Russia

cpshenichny@yandex.ru

Abstract. Chemistry, despite its high level of formalization, benefits from implementation of knowledge engineering tools. "Static" (or object-based) methods have been successfully used in this science, but the character of chemical knowledge urges one to look also for "dynamic" (event-based) methods, especially in experimental and industrial domains. Still, quite a work is needed to make the application of event-based methods in chemistry as perfect and correct as that of object-based ones, and adjustment of the said methods to this science may considerably contribute to the theory that underlies them. In particular, new solutions have been found at testing the method of event bush by chemical tasks. These solutions may optimize the method for use in a wide range of fields.

Keywords. Chemistry, knowledge engineering, dynamic knowledge, event bush method, experiment.

1 Introduction

Since the genius discovery of Dmitry Mendeleev expressed in his periodic table of elements, chemistry became one of the best-organized fields of knowledge the humankind ever had. Nevertheless, abundance and diversity of combinations of "allowed" bonds between the elements, especially in organic, bio- and geochemistry, make this field, despite its internal regularity, rather loose and hard-to-span. This claims for application of special methods of organization of knowledge, and such methods have been successfully applied in chemistry.

The most extended and powerful tools of knowledge organization developed for chemistry and related fields (biochemistry, medicine) are Chemical Entities of Biological Interest (CHEBI) ontology [5], nomenclatures of compounds produced by International Union of Pure and Applied Chemistry [3], chemical divisions of Chem-ID Plus, Medical Subject Headings [10] and some other search systems. In some of them not only lingual but also visual means of representation (structural formulae of compounds) are implemented (e.g., in CHEBI [4]) that is an unusual solution for the ontology design.

Nevertheless, these developments generally fail to encompass one important feature of chemical knowledge – its dynamic character. Indeed, nomenclature of elements or compounds is, in fact, only an introduction to understanding of possible chemical reactions, this well-ordered miracle of transformation of one substance into another. Verbal and graphic explication of this order is useful for understanding of various branches of chemistry, for planning the experiments and, of course, for industrial applications. This is why there were a number of attempts of using the event-based methods in chemical or related issues. Event trees and Bayesian networks are involved to model hazards and disorders at chemical factories [1], flowcharts (sensu stricto or sensu lato, sometimes quite informally), to describe experiments [2; 6] and even to perform classification of compounds [7] – i.e., to approach a "static" task from "dynamic" side.

Still, these methods, as was argued by Pshenichny and Mouromtsev [9], contrary to the object-based ("static") ones, need better rethink and formalization of grammar of event description in the nodes. In such a formalized domain as chemistry, this "under-formalization" of knowledge engineering approaches becomes especially evident. One method of dynamic knowledge engineering, the event bush, has got the most evolved verbal grammar and related structural rules of event combination [8]. In our paper an opportunity of formalization of record of events, processes and scenarios in chemistry by means of the event bush method will be examined. For this, first there will be considered a trivial task from inorganic chemistry, and then, it will be transformed into a hypothetic experimental/technological application, which also will be modeled by event bush.

2 An Example of Formalization of Simple Chemical Reaction

For the beginning, one of the simplest and best-known chemical reactions was considered, that of acid and alkali with formation of salt and water – e.g., $HCl+NaOH=NaCL+H_2O$; $H_2SO_4+CaO=CaSO_4+H_2O$.

On the one hand, the case looks obvious for application of the event bush method. For this, we shall conceptualize this equation as a possible scenario in some environment of directed alternative changes, in which we define the "key players" (primary internal, or ia events) as primary, non-unique inputs, and external "actors" (primary external, or ib events) that may put some constraints on the behavior of "key players" [8]. This "distribution of roles" looks straightforward in the considered case – one of the participating compounds is "key player", while the other, "external actor".

However, on the other hand, there is a semantic complication in this seemingly trivial case. Both the alkali ad acid enter reaction symmetrically, and there is no ground to prefer one as "key player". Be acid affected by alkali or alkali affected by acid, the result (salt+water) would be the same. Still, this complication gives us a methodologically beautiful opportunity to compose two event bushes, in which the ia and ib events change places. The result is shown in Fig. 1 a,b.

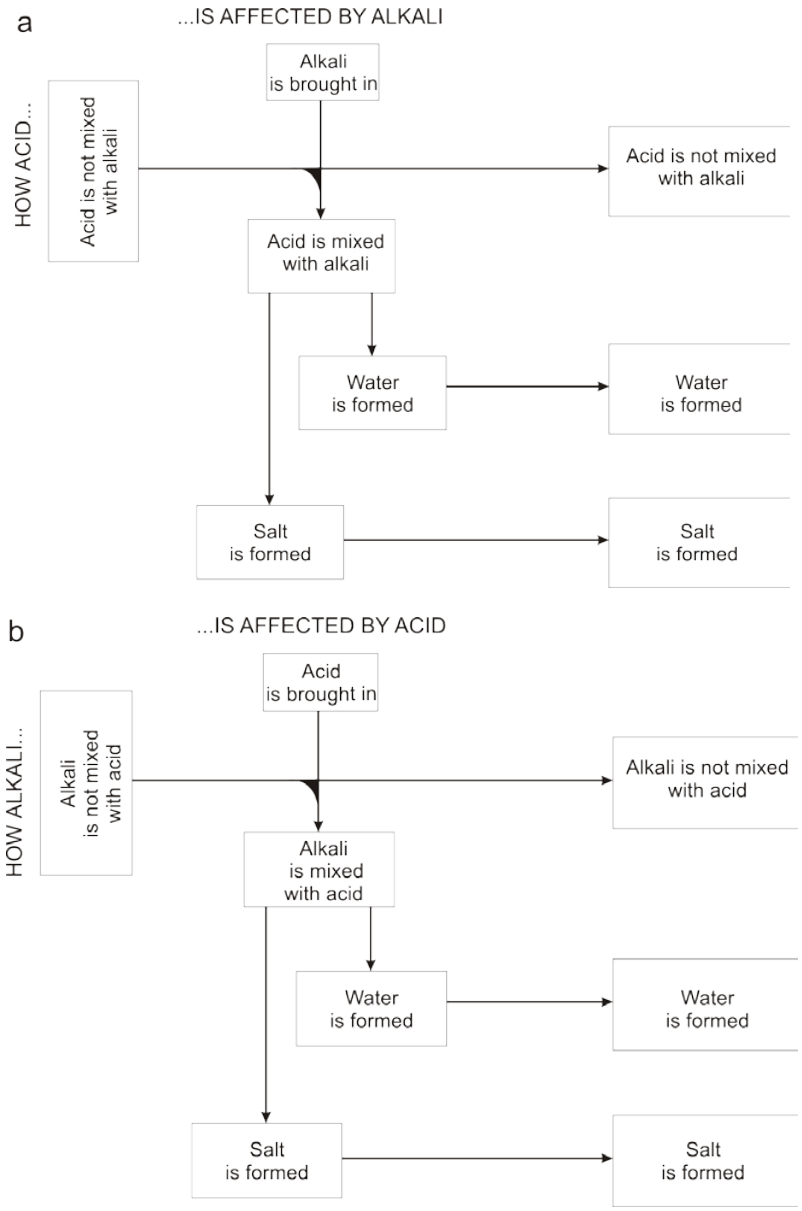


Fig. 1. Two alternative event bushes describing similar reaction of acid and alkali. See comments in the text.

In accordance with the character of the reaction, the two bushes look quite symmetrical. In both of them two incompatible scenarios are seen: one, the mandatory

"nothing happens" scenario (if compounds are not brought together – in the event bush semantics, either alkali is not added to acid, or vice versa), and the other scenario that depicts the reaction resulting in simultaneous formation of salt and water. As a methodological experience, one may conclude that dealing with a case that two events happen simultaneously, equally influencing each other and symmetrically determining the future course of events, a couple of event bushes with symmetrical structure has to be expected as shown in Fig. 1 a,b. Because of similarity of consequences, it can be postulated that "Acid is mixed with alkali" is equivalent to "Alkali is mixed with acid". (Though, the meaning of equivalence so far is understood here rather informally; there is not enough ground to appeal to definition of equivalence used in any existing formal system, e.g., in classical logic, because the event bush has not been entirely interpreted in terms of any of such system.)

One may suppose that considering a reaction involving three or more compounds may represent a problem because the event bush semantics implies only two types of primary events, primary internal (ia) and primary external (ib), and this division is related to the binary subject-predicate structure of statements representing events in the event bush [8]. Still, it looks unlikely that three or more agents interact with each other exactly in one time, and if not, there should be one-to-one collisions, and the whole reaction can be represented by successive or parallel *couple interaction*s, i.e., be well modeled by event bush (or a pair of bushes).

The above example shows an ability of event bush to cope at least with some basic issues of pure chemistry. Below an applied issue will be considered.

3 Experimental and Technological Application

To address an applied chemical issue, suppose a very simple example of experiment or production – a tank filled with two liquids (fluids) divided by an impermeable screen. The screen is removed; fluids contact each other and mix (Fig. 2).

This simple case represents a purely mechanical process and may be remarkable only by the use of one more, optional connective of the event bush, the conflux (see the bottom of Fig. 2). However, along with the "pure-chemical" case depicted in Fig. 1, this is just an "introduction" to an "applied chemical" case. Suppose that one fluid in a tank is acid and the other, alkali. What happens then is modeled by the event bush in Fig. 3.

To build this bush, "Fluid A" in the bush from Fig. 2 was changed to "Acid", and "Fluid b", to "Alkali". In all the rest, the upper part of the resulting bush repeats the bush in Fig. 2. Hence, the new bush was derived from the previous one by substitution of two subjects. In other words, given this substitution, the bush in Fig. 2 *is the rule* for construction of the upper part of the new bush. However, when replacement reaches the event "A mixture of fluid A and fluid B is formed in the tank", this results in event "A mixture of acid and alkali is formed in the tank" of the new bush. From this point, based on the meaning of the considered events, one may postulate that the

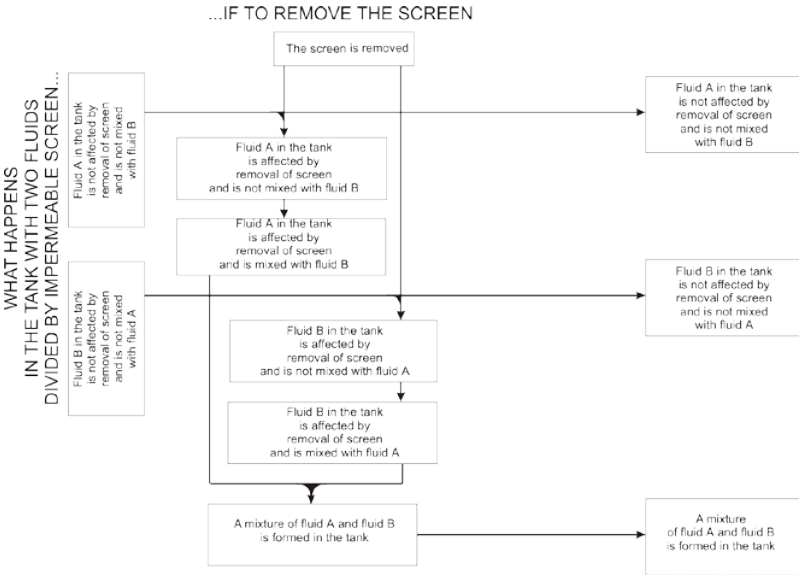


Fig. 2. An event bush describing the behavior of two fluids in a tank (see comments in the text)

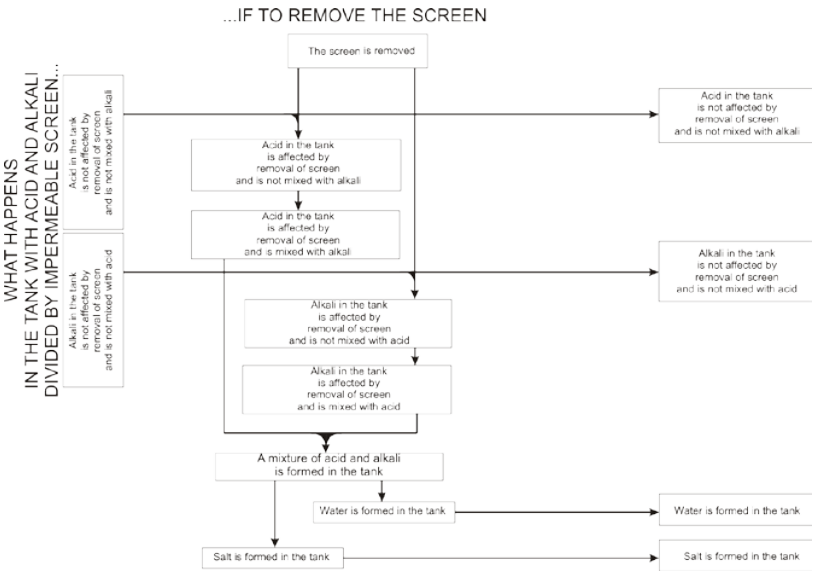


Fig. 3. An event bush describing the behavior of acid and alkali in a tank (see comments in the text)

Fig. 2 bush does not apply as a rule anymore for the newly constructed bush. Instead, it looks reasonable to postulate that the event "A mixture of acid and alkali is formed in the tank" is a kind of (or a particular case of) event "Acid is mixed with alkali" or its equivalent "Alkali is mixed with acid". Then, the rest of the bush will be composed based on the corresponding part of the Fig. 1/Fig. 2 bushes. Thus, there is no tertiary event that would correspond to "A mixture of fluid A and fluid B is formed in the tank" of Fig. 3 bush, but instead there are secondary statements that are particular cases of "Water is formed" and "Salt is formed" Fig. 1/Fig. 2 bushes – "Water is formed in the tank" and "Salt is formed in the tank", correspondingly, and the same-formulated tertiary results.

Despite the triviality of the case, it demonstrates an important methodological novelty. Some event bushes serve as the *rules of composition* for another bush.

4 Discussion

The rules of composition of one event bush based on others need to be formalized to become independent of meaning of particular events. This seems to become feasible with complete formalization of event description grammar and algorithmization of building the event bush. Nevertheless, what can be definitely said now is that having a number of event bushes constructed, one can *obtain new knowledge* combining them, binding them with additional axioms and thus constructing new bushes. (Another issue is how well this knowledge would be supported by data.)

One way of building a bush based on another bush is specification of events and substitution of genus by differentia in subjects or predicates of some events, e.g., "Fluid A" to "Acid" or, possibly, "Acid" to "Formic acid" in Figs. 1 a,b. If to continue this approach and descend down to instances, e.g., to "Formic acid sample no. 49276" instead of "Formic acid", the event bush may be transformed into a data-storing facility. Also, attributing quantitative values to the events of the bush and attributing computational sense to its connectives, one may create a tool for computation of chemical reactions or physical-chemical or technological computation [8].

Theoretical findings made at adjustment of the event bush method to pure-chemical and applied chemical tasks (a couple of equivalent bushes and understanding of event bush as a rule for composition of another bush) emerged at the very beginning of application of this method in chemistry. It looks highly probable that further research in this direction will bring the results that will enrich the theory of event bushes and serve in many other fields to organize existing knowledge and, perhaps yet more importantly, obtain new one.

5 Conclusions

1. A methodological novelty brought by testing of the method of event bush by modeling simple chemical reactions is an opportunity to construct a couple of equivalent event bushes that model similar environment equally well but should be

considered in pair to reflect the observed symmetry of primary internal and primary external events.

2. Putting a primitive experimental/technological task in terms of event bush has revealed an important opportunity to use one bush as a rule of composition of another and therefore obtain new knowledge combining existing event bushes.

3. At present, building new event bushes based on existing ones is performed largely by intuition and for trivial tasks; formalization of this procedure will open wide opportunities for dynamic knowledge engineering in various fields.

References

1. Azhdari, M., and Mehranbod, N.: Application of Bayesian belief networks to fault detection and diagnosis of industrial processes, in: International Conference on Chemistry and Chemical Engineering (ICCCE), Proceedings, 1-3 Aug. 2010, 92-96 (2010).

2. College of Science and Technology, Armstrong Atlantic State University: Chemistry Laboratory Example Flow Chart, http://chemistry.armstrong.edu/nivens/Chem2300/flowchart.pdf (2012)

3. Compendium of Chemical Terminology, Version 2.3.2. International Union of Pure and Applied Chemistry: 2012

4. Ennis, M.: ChEBI A Dictionary of Chemical Entities with an Associated Ontology, in: SOFG-2 Proceedings, Philadelphia, October 23-26 (2004).

5. Hastings, J., Magka, D., Batchelor, C., Duan, L., Stevens, R., Ennis, M., and Steinbeck, C.: Structure-based classification and ontology in chemistry. Journal of Cheminformatics, 4, 8 (2012).

6. Kuwata, K.: Analytical Chemistry Laboratory Notebook Guidelines, http://www.macalester.edu/~kuwata/Classes/2004-05/chem%20222/chem%20222%20notebook%20directions%202005.pdf (2004-2005)

7. Marr, K.: Chemical Nomenclature Flowchart, in: Chemistry 161, http://www.instruction.greenriver.edu/kmarr/Chem%20161/Chem%20161%20ALEs/POGIL%20ALEs_F2008/Chap%202%20POGIL%20ALEs_F2009/9x_Nom%20Chart_Elements_Ions%20_Practice_Ans%20Key_Ch2_F2010.pdf (2001)

8. Pshenichny, C.A., and Kanzheleva, O.M.: Theoretical foundations of the event bush method. In Societal Challenges and Geoinformatics, GSA Special Paper 482, Sinha, K, Gundersen, L., Jackson, J., and Arctur, D. (Eds.), 139-165 (2011).

9. Pshenichny, C.A., and Mouromtsev, D.I.: Representation of the Event Bush Approach in Terms of Directed Hypergraphs, in: ICCS Proceedings, International Conference on Conceptual Structures. Pfeifer, D., and Ignatov, D., (Eds). Mumbai, January 2013, in press (2013).

10. U.S. National Library of Medicine: Medical Subject Headings, http://www.nlm.nih.gov/mesh/ (1999-2013).

Dynamic Information Model for Oceanographic Data Representation

Natalia A. Zhukova[1], Dmitry I. Ignatov[2], OksanaV. Smirnova[1]

[1]Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg, Russia
nazhukova@mail.ru, sov@oogis.ru
[2]National Research University Higher School of Economics, Moscow, Russia

Abstract. This paper treats issues of dynamic information model for oceanographic data representation construction. Proposed model includes three submodels – statistical model of data description, logical model of data relation description, model description of processes of change of water environment parameters state. As well statistical and intellectual methods used for automation of data processing and analysis are presented. Use of these methods will allow reducing processing time, to provide possibility of adaptive dynamic data processing generation, to improve processing which assumes data handling not at the level of measured values, but at the level of knowledge about measurements, parameters, and their relationship and also knowledge about subject domain.

Keywords. dynamic information model, intelligent processing of oceanographic data, geoscience, data processing.

1 Introduction

At the present time interest to problems relating to research of environment conditions significantly increased. It is, first of all, due to changes in the atmosphere, ocean and earth's surface caused by different factors. Secondly, methods of data processing and analysis, that were developed, are oriented on use by subject domain experts. Generally data processing and analysis are performed by hand using special tools. Today three are three main problems – first, the low speed and quality of newly received data acquirement, secondly, complexity and low speed of data processing in delayed mode, thirdly, complexity of the task solution of forecasting water environment state. At the stage of operative data processing preliminary estimation of data quality is performed. Quality rating is held with the use of test set specialized for different data sources and regions and takes about a day.

The most difficult operations are operations of analysis in the delayed mode. The procedure of delayed data processing provides removal of noise and outliers, that don't differ much from measurements, and restoring of missing values, calculation of

offsets, exposure of trends, comparison with statistical data for detection of data correctness.

Experts have to analyze in details data when performing processing in the delayed mode taking into account all earlier received data on the area of interest, data received using intended and similar data sources, and also knowledge of physical features of the environment of the studied region. Complexity of problems of the delayed processing constantly increases as the volume of data which must to be processed increases. So measurements or result of their processing are available to end users on the average in half a year after receiving measurements. Also, part of errors is removed well after, and the general time of identification and removal of errors can take about two years.

Users of oceanographic systems (for example, tools for hydroacoustic monitoring of the water environment) have to deal with all these problems. For the analysis of oceanographic data ready-made products of the analysis are used that are usually updated two times in a year. Thus access to operational data isn't provided. It leads to decrease in accuracy of estimate of water environment state and, respectively, decrease in operating benefits of hydroacoustic tools.

The considered problems can be effectively solved at the expense of use of dynamic information model for oceanographic data representation that is reflective to actual state of water environment and also state of subject domain objects. The basis of proposed dynamic information model is a set of three models. It is statistical model of data description, logical model of the data relations description, model of the description of processes of change of a water environment parameters states. The dynamic information model is constructed on the basis of set of data mining methods.

2 Description of Oceanographic Data

For 30 years the basic source of ocean data was data received from oceanographic stations and mooring buoy station. Total amount of data made was about 500 measurements per day. Argo project [3] was started in 2000. The target number of Argo buoys was 3000. Currently general number of buoys are 122, general number of measurements are 109050. Number of oceanographic stations, bathythermospheres, buoys constantly increases. From all sources about 2000 measurements are received each day. At the present time total number of stations is about 12 million. Total amount of the available data contains 14 million of temperature profiles and 5 million of salinity profiles. Each profile represents set that contains time, earth coordinates, depth level and related measurements. Figs. 1-3 illustrate examples of temperature and salinity measurements.

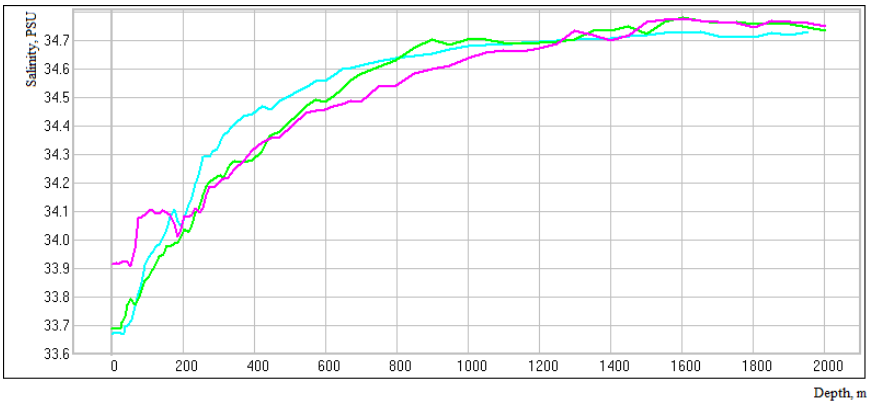


Fig. 1a. Salinity measurements.

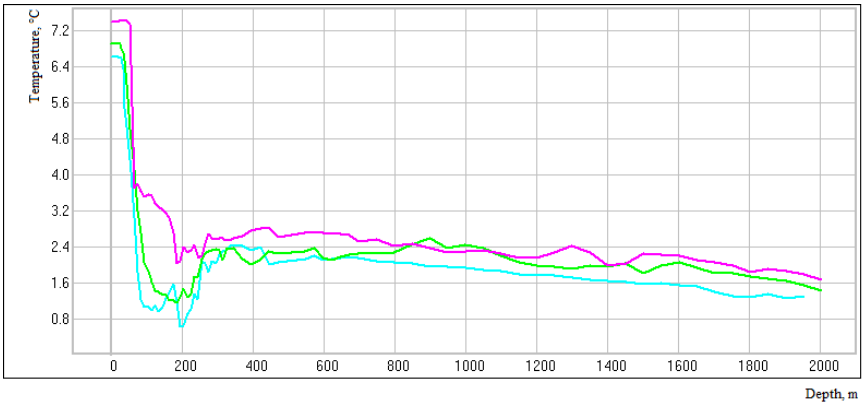


Fig. 1b. Temperature measurements.

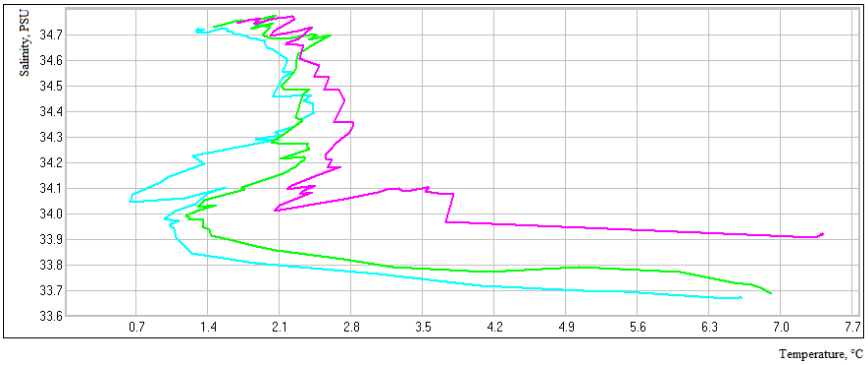


Fig. 1c. Temperature VS salinity.

Fig. 1. Measurements from Argo buoys (region is (49.5-50.5° S, 37.6-38.2° W)).

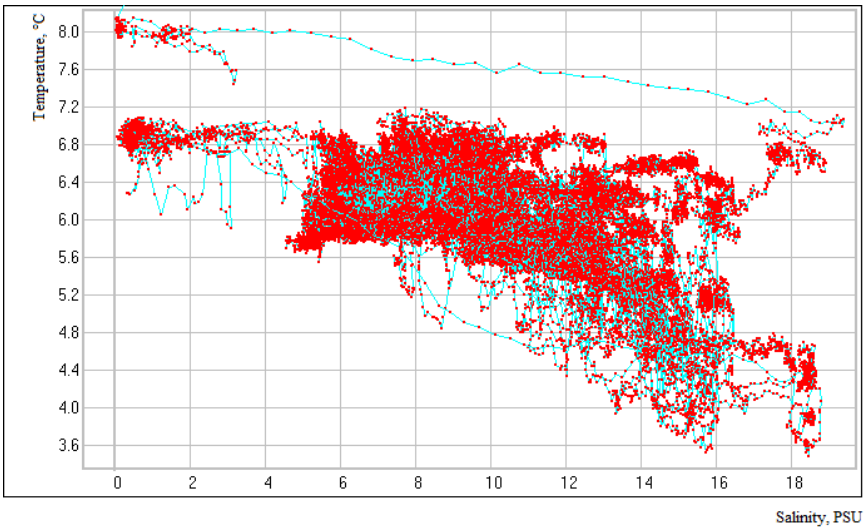


Fig. 2a. Temperature VS salinity.

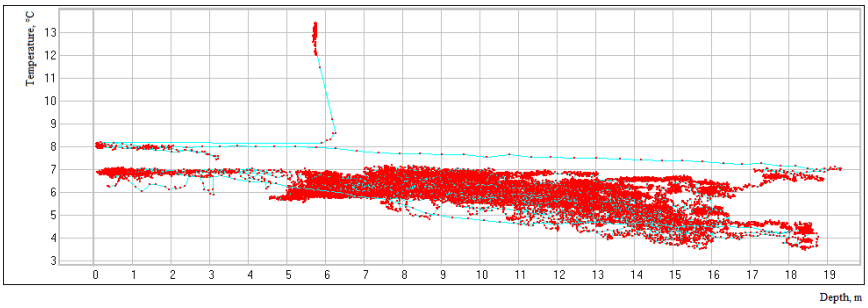


Fig. 2b. Temperature VS depth.

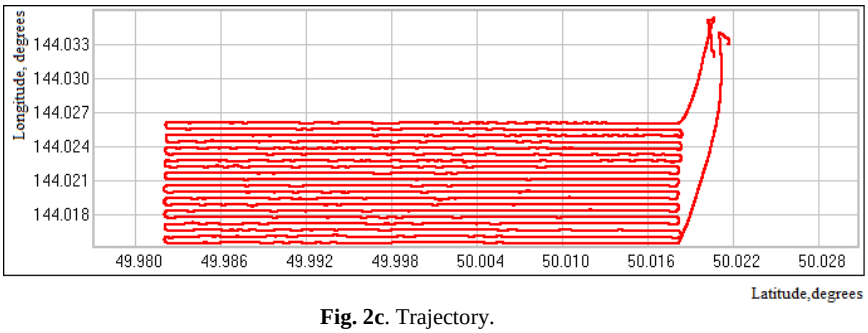


Fig. 2c. Trajectory.

Fig. 2. Measurements from autonomous underwater vehicle (region is (49.9-50.02° S, 144.0-144.033° W), time period is October).

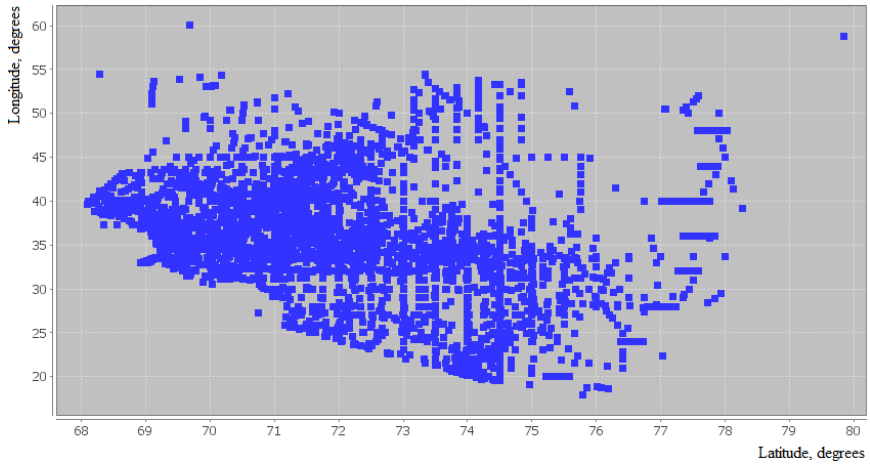


Fig. 3a. Latitude VS longitude (depth is (0-5).

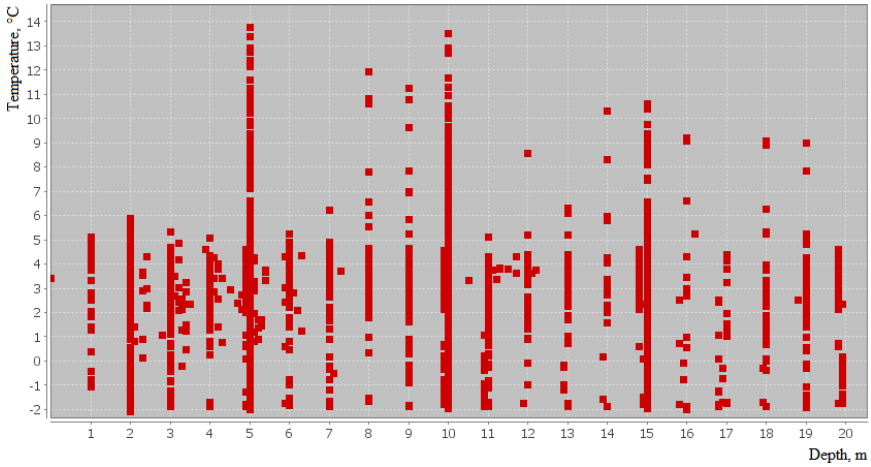


Fig. 3b. Depth VS Temperature.

Fig. 3. Measurements from various oceanographic stations (region is (60-80° N, 15-60° E), time period is January).

Analysis of data showed that measurements are not regular both by time and coordinates with the exception of data received from fixed stations. In additional data and its statistical characteristics for different regions are far from being similar and need special solutions for their processing. As well, processing of measurements assumes specialized methods, particular at the stage of data quality estimation, specific for each type of source.

Measurements have following particular characteristics:

- measurements are time series with different behavior. That is because they are received in different regions using different measuring tools. A set of external factors influence strongly on received values, for example, the seasonal phenomena, state of water environment of contiguous area.
- data contains considerable number of error values, for example, noise, outliers, gaps and also offsets and trends due to errors of measuring tools. Measurements on each data source and on each region demand application of specialized methods of processing. The majority of them requests participation in process of the expert.

Problem solution of automatic choice of methods and definition of their parameters is carried out by means of the use of adaptive approaches to data processing based on domain knowledge and statistical data.

3 Description of Dynamic Information Model for Oceanographic Data

Dynamic information model is integrated model of oceanographic data description developed on the basis of historical data and expert knowledge of subject domain. It provides actual data corresponding to environment and objects settings. Components of this model, which are based on [4, 5, 6], are [7, 8]:

- statistical model of data description is used for formalized description of separate measurements and their set, and also knowledge about the measurements received as a result of their processing. The following types of data and knowledge representation are used: the initial measurements representation including initial measurement representation model, results of data harmonization representation, including models of structural measurements representation, representation of results of data integration, including models of semantic representation of measurements;
- logical model of data relation description is a set of models that includes: models of representation of data integration results, including models of representation of multidimensional measurements, models of the qualitative and quantitative data description, representation of results of data fusion, including models of heterogeneous data combined representation;
- model of the description of processes of change of water environment parameters state represents relations between different processes on quantity and quality levels.

Dynamic information model for oceanographic data allows solving three main problems:

1. provides representation of actual information on various subject domain objects and states water environment parameters at a given moment of time and given point in space and possibility of operative improvement of information as a result of processing of the received data;

2. provides the short-time forecast of a state of basic parameters of the environment taking into account available data, knowledge and factors, that impact on state of parameters;
3. provides information on data relation and dynamics of the parameters change.

Primary properties of dynamic information model are:

- model is multilevel in the context of information content, it contains information of various levels - from initial data to knowledge about processes;
- model is multilevel and hierarchic and reflects the structure of subject domain– from separate measurements and group of measurements to measurements of separate regions;
- model is multidimensional (with different granularity);
- model is capable to accumulate all previously gathered data and knowledge;
- model is capable to provide rating and accounting of external factors, that influence directly or indirectly on state of the environment.

Harmonization, integration and fusion data [9, 10] and also statistical analysis and data mining are key technologies that are used in the dynamic model.

Data harmonization suppose definition of main concepts and their relationship on the corresponding subject domains and/or responsibility spheres. The general procedure of data integration assumes: an assessment of data quality from each source on the basis of specialized set of tests; search and exclusion of duplicating values; statistical data processing of each set of measurements, including denoising, removing outliers, identification of trends, filling gaps; interpolation of data. Data fusion is defined as process of data combination from various sources which allow to receive information of new quality and reduce its size. Statistical analysis and data mining provide task solution of system data processing and knowledge acquisition from data.

4 Description of Data Mining Technology

The general method of multidimensional measurements analysis using data mining methods is given in Fig. 4. Proposed stages are general and depending on data type and solved task stages can be skipped.

Stage 1. « Structure analysis». For initial data, that is a structured binary stream, that contains measurements, task of stream structure validation is solved.

Stage 2. «Measurements extraction». Measurements extraction assumes parameter measurement extraction from data stream according to its description.

Stage 3. «Definition of measurement types». For each measurement, received on the stage of measurements extraction, its type is defined. Parameters, which possess priori formed set of properties, refer to one type. The constant and spinner can be examples of measurement types.

Stage 4. «Data preprocessing». Cleaning measurements from noise or outliers, exclusion of trends, filling missing value is implemented on the stage of data preprocessing. In additional, statistical analysis of measurements is fulfilled, for example, statistical analysis of distribution parameters, regression analysis, spectrum analysis.

Stage 5. «Data segmentation». The stage assumes segmentation of time series, so that each segment has a defined set of constant properties. Segmentation can be realized by experts or using segmentation algorithms. Segments and their characteristics are saved in model. When new data is received, it is segmented taking into account results of segmentation of historical data.

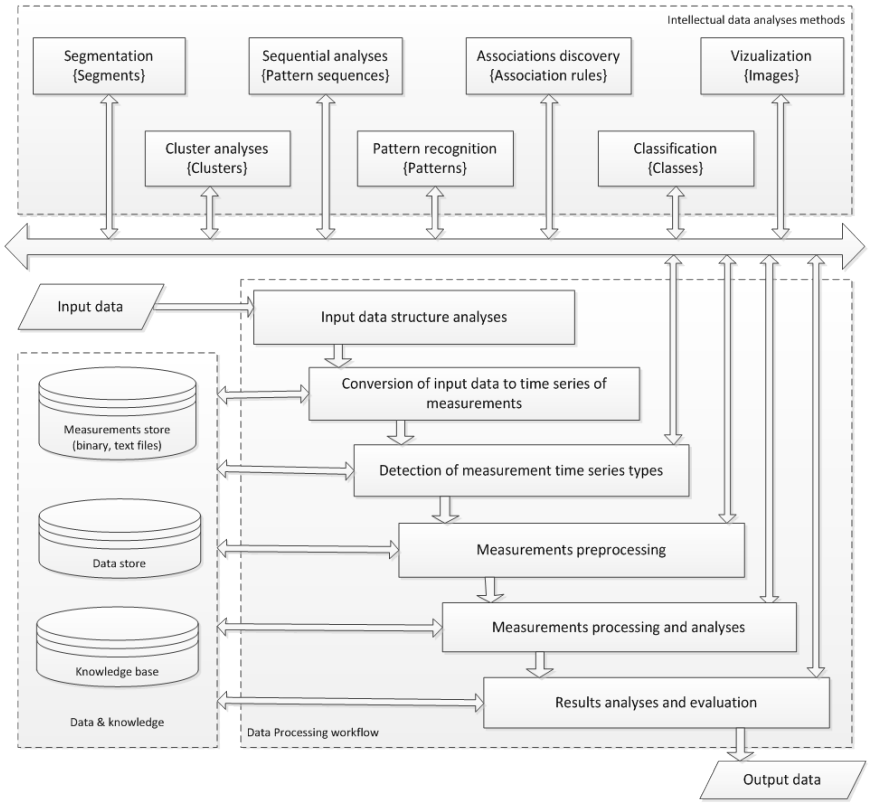


Fig. 4. General methods of multidemensional measurements analysis using methods of data mining.

Stage 6. «Cluster analysis». Clusterization problem consists of detection and description of confluence areas in analyzed space i.e. clusters are defined so, that distance between instances of one cluster is minimal and distances between instances of different clusters was maximal. Procedures for distances calculation are defined using specified criteria. When clustering time series first segmentation is made. Application of cluster analysis algorithms to time series allows revealing a set of possible time series states.

Stage 7. «Sequential analysis». This stage supposes searching time dependencies in sequence of segments. Time dependencies are represented in the form of a pattern

sequences. Formed patterns are saved in model. When analyzing new data, match of new data to patterns is checked.

Stage 8. «Association analysis». The stage assumes search of association dependencies in interval and qualitative data in the form of association rules. The rules are mined in historical data and then they are located in knowledge base. Discovered rules are applied for analysis of new data.

Stage 9. «Pattern recognition». The stage is intended for generation of measurement pattern on the basis of single-type measurement. Recognition of new data is realized by comparing new data and patterns.

Stage 10. «Visualizing results». When working with historical data analysis of initial data and results of analyses at different stages are visualized. When analyzing new data discovered mismatches are visualized.

Stage 11. «Obtained results analysis». This stage supposes representation of data processing results, oriented on expert use. It assumes usage of cognitive graphics methods and other visualization tools. At this stage formation or extending of knowledge base is realized.

Automation and adaptation of data mining processes and analysis of multidimensional measurements is performed by means of use of exploratory analysis and mechanisms of processing control. Procedures of prospecting analysis, that allow to receive priori estimates of data. According to estimations and using classification of measurement type and rules for data and knowledge representation of different types effective form of measurements and knowledge representation can be chosen and appropriate processing methods can be used. Mechanism of processing control is one of the central element in data processing and analysis systems. It provides data processing processes construction and correction. Mechanism of processing control is described in [11, 12].

5 Presentation of Dynamic Information Model in Intelligent Geoinformation System

Dynamic information model for representation of oceanographic data is realized under system of lighting situation. It is oriented on solution of wide range of problems, for example, search, detection, classification, definition of different objects parameters and also solution of hydroacoustical problems. Description of architecture of intelligent geoinformation system (IGIS) gives in [13]. Figs. 5-6 display examples of processed data and result of regular data grid construction on the basis of dynamic information model.

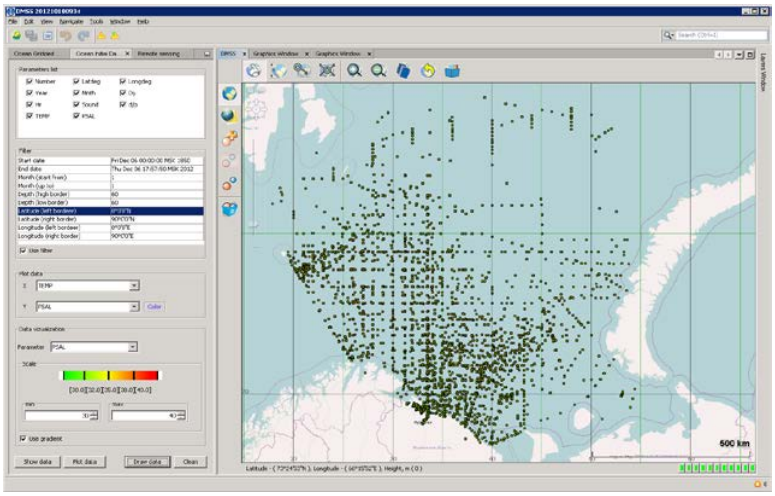


Fig. 5. Visualization of processed oceanographic data in IGIS.

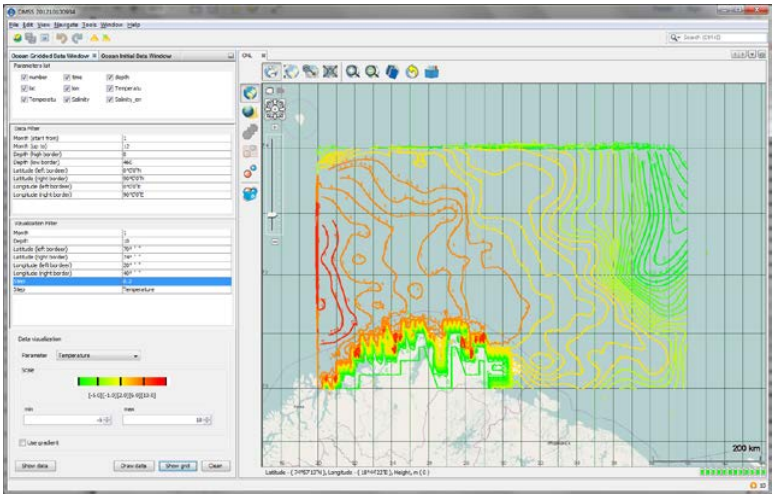


Fig. 6. Visualization of gridded data in IGIS.

6 Case Study

The dynamic information model of ocean data representation was constructed on the basis of data received from Arctic region during the period from 1876 up to now [13, 14]. Temperature and water salinity of Arctic region were measured at depths from 0 to 460 meters. Total number of performed measurements is about two million. Data-

base of measurements is made and provided by the Arctic and Antarctic research institute on a grant of Office of Naval Research #62909-12-1-013 ("Decision Making Support System for Arctic Exploration, Monitoring and Governance"). In Fig. 7 temperature and water salinity values distribution and distribution of gathered data by years and depths is shown.

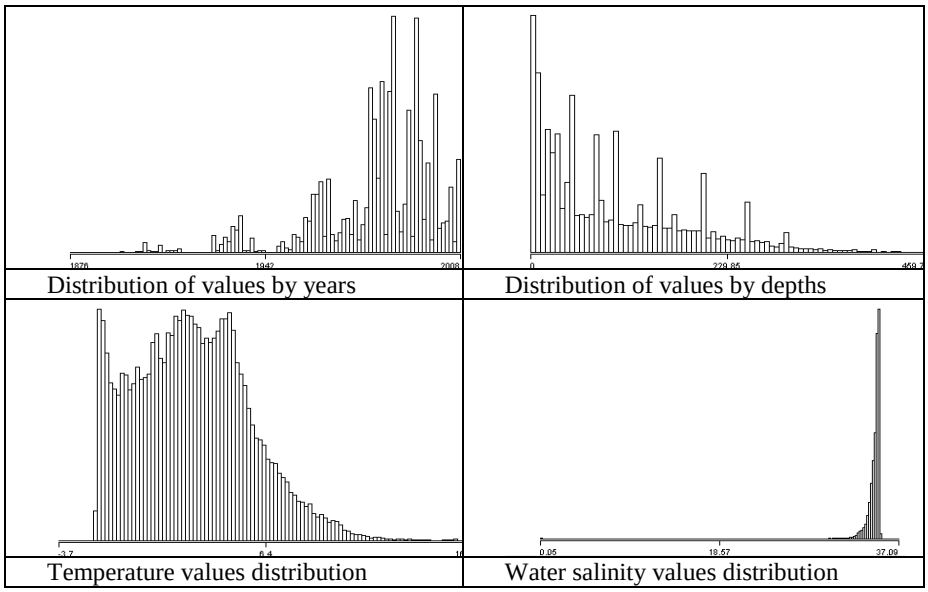


Fig. 7. Temperature and water salinity values distribution by years and by depths.

Example of application of data mining methods for solving task of operational data assessment obtained from external sources for the propose of decision making if it can be used at the next processing stages, particularly, for recalculating nodes of regular grid is given.

The task solution of operative data assessment is founded on comparison of received data with historical data of the same region at similar time intervals. As time line months in which measurements were received were considered. One of the most complex tasks is detection of stable regions in which values of analyzed parameters differ slightly. Task of region detection was solved using methods of cluster analysis. Below the description of procedure of region detection based on analyses of data received in various years in July is provided. As algorithm of cluster analysis SimpleKMeans algorithm was used, number of clusters was selected using estimation of result clusters compactness.

Step 1. Cluster analysis of initial data: time interval – from 1870 to 2008, time period – July, range of depths – from 0 to 460 meters, elements of feature space – latitude, longitude and depth of measurements, year of measurements conduction, values of temperatures and salinities. Results of cluster analysis are shown in Fig. 8, descrip-

tion of clusters is given in the Table 1. Borders of clusters take place at depths of 40 and 120 meters which is equivalent to border of water layers.

Table 1. Description of cluster centers developed for initial sample.

Attribute	1	2	3	4	5	6	7	8	9	10
Latitude, degrees	76.2	71.5	71.1	70.8	70.3	71.3	72.4	76.3	76.4	77.2
Longitude degrees	27.4	33.9	32.2	32.8	50.2	34.5	31.4	28.1	30.7	51.5
Depth, meter	77.4	110.5	28.2	12.3	22.7	98.8	246.9	17.4	179.4	58.5
Tempera-ture, celsius	0.6	3.2	5.9	6.6	3.4	2.8	2.3	1.8	0.7	-0.5
Salinity, PSU	34.6	34.7	34.1	33.7	32.7	34.7	34.9	34.1	34.8	34.4

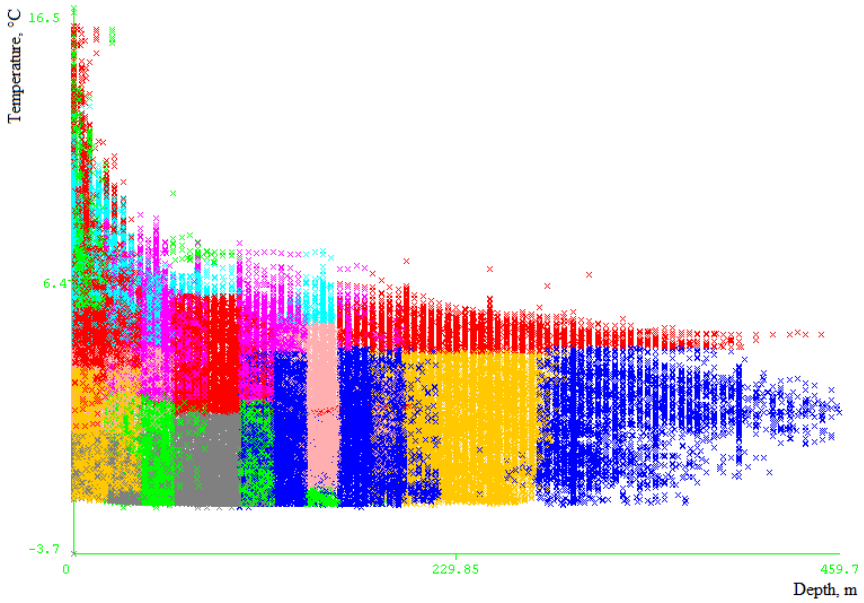


Fig. 8. Results of cluster analysis of initial samples.

Step 2. Cluster analysis of the data measured at depths of 0 - 40 meters. Results of the cluster analysis are shown in Fig. 9. Total number of the clusters are 5. At depth around 20 meters clear boundary of clusters is observed. It means that further data partitioning by parameter "depth" is to be done.

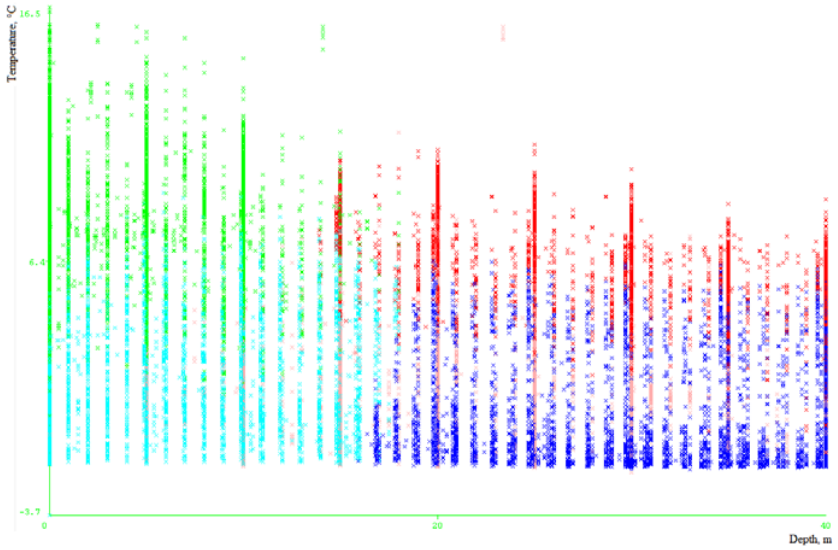


Fig. 9. Results of cluster analysis of data for depth 0-40 meters.

Step 3. Cluster analysis of data measured at depths of 0 - 20 meters. Results of the cluster analysis (Fig. 10) show that further data partitioning by parameter "depth" isn't expedient. However, clear clusters of measurements can be observed in the space of latitude and longitude (Fig.11) features. As boundary value latitude of 74 degrees is considered.

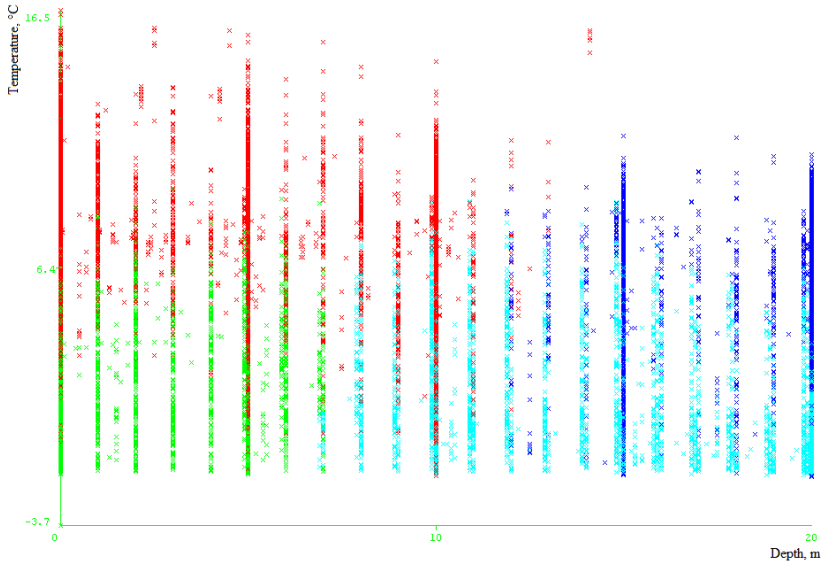


Fig. 10. Results of cluster analysis of data for depth 0-20 meters (Depth VS temperature).

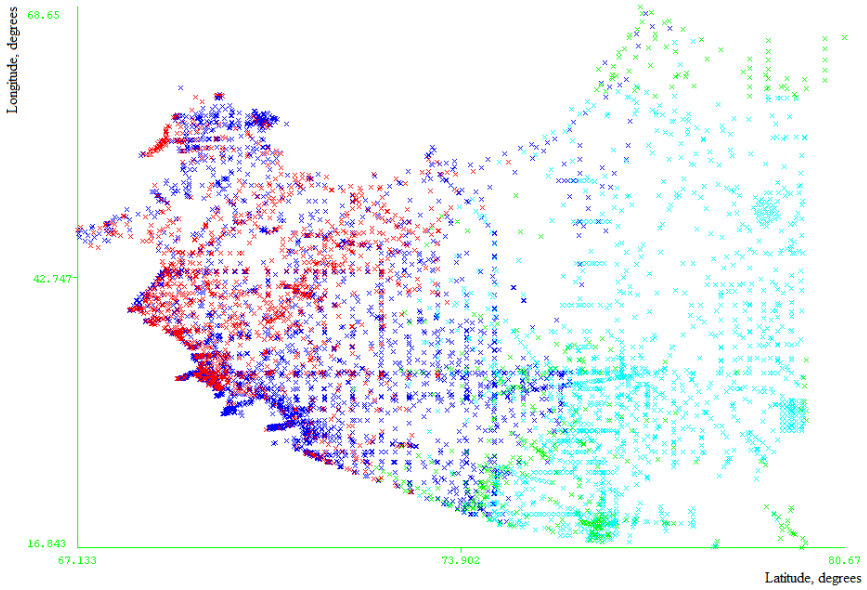


Fig. 11. Results of cluster analysis of data for depth 0-20 meters (Latitude VS longitude).

Step 4. Results of cluster analysis of data for depth from 0 to 10 meters and latitude more than 74 degrees is shown in Fig. 10. Further decomposition of data was done by parameter "longitude", for measurements with value of latitude more than 42 degrees.

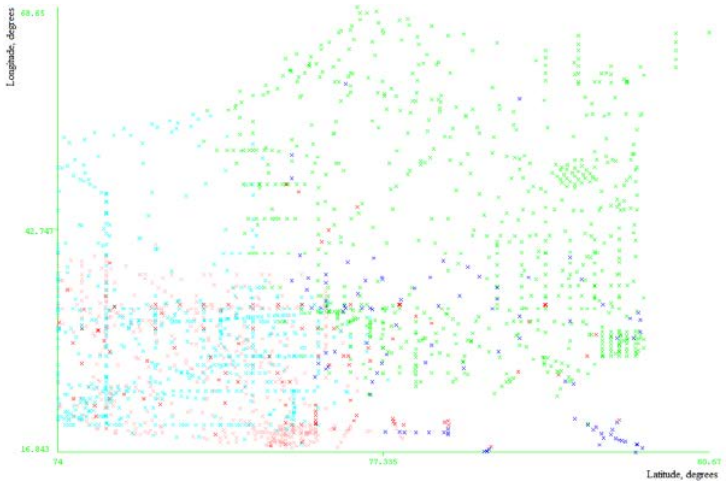


Fig. 12. Results of cluster analysis of data for depth 0-10 meters, latitude 74-81 degrees.

Step 5. Results of cluster analysis of data for depth from 0 to 10 meters and value of latitude more than 74 degrees and longitude more than 42 degrees are given in Fig. 13. At this stage data non-crossing clusters are formed.

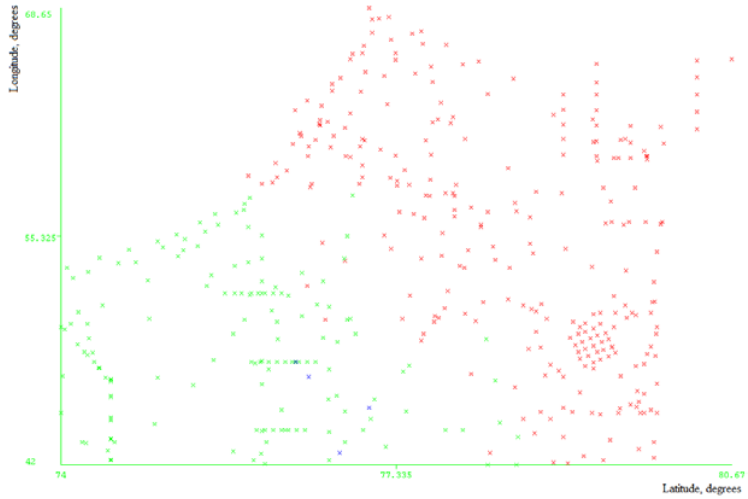


Fig. 13. Results of cluster analysis of data for depth 0-10 meters, latitude 74-81 degrees, longitude 42-69 degrees.

In the similar way all historical data on Arctic region was analyzed. All data space was decomposed on set of stable regions.

Results of cluster analyses were interpreted by specialists from Arctic and Antarctic Research Institute (Saint-Petersburg, Russia). Clusters border at depth of 20 meters corresponds to a wave-mixing zone. The zone exists during the time when there is no ice. Borders of clusters for depth of 0-20 meters are not quite clear because in July seasonal thermowedge is destroyed. Spatial distribution of data show zone of the Norwegian current, borders of distribution of Atlantic waters in Barents Sea.

7 Conclusion

The paper illustrates the dynamic information model for oceanographic data representation based on application of data mining methods and intelligent GIS technologies. Proposed model allowed to decrease processing time both in operational and delayed mode due to use of automated methods of data analyses, such as cluster analyses. That is important for different monitoring systems of water environment.

The further direction of researches is connected with application of biclustering and triclustering methods to oceanographic data. These methods are nowadays widely used in various spheres [1, 2, 3]. That allows take into account not only measurements but also time and location where measurements were received, so it can be expected that the rate of cluster compactness will increase.

Reference

1. Gnatyshak, D., Ignatov, D.I., Semenov, A. and Poelmans, J.: Gaining Insight in Social Networks with Biclustering and Triclustering. In.: Aseeva Natalia, Babkin Eduard, Kozyrev Oleg (eds.) Perspectives in Business Informatics Research, Lecture Notes in Business Information Processing. Volume 128, Part 4 (2012) 162-171
2. Ignatov, D., Poelmans, J., Zaharchuk, V.: Recommender System Based on Algorithm of Bicluster Analysis RecBi. In CEUR Workshop proceedings, CDUD'11 – Concept Discovery in Unstructured Data. Volume 757 (2011), 122-126
3. Ignatov, D.I., Kuznetsov, S.O., Magizov, R.A. and Zhukov, L.E.: From Triconcepts to Triclusters. In: Kuznetsov et al. (eds.) RSFDGrC 2011, LNCS/LNAI, vol. 6743/2011, pp. 257-264. Springer-Verlag Berlin, Heidelberg (2011)
4. The International Argo Project Homepage, http://www.argo.net/
5. W3G Geospatial Ontologies, http://www.w3.org/2005/Incubator/geo/XGR-geo-ont-20071023/
6. OGC Standards and Supporting Documents, http://www.opengeospatial.org/standards
7. Llinas, J., Bowman, C., Rogova, G., Steinberg, A., Waltz, E., White, F.: Revisiting the JDL data fusion model II. In: Proceedings of the Seventh International Conference on Information Fusion, Stockholm, Sweden (2004)
8. Liss, A.R., Zhukova, N.A.: Software System for Processing On-Line Information of Complex Dynamic Objects. In: Proceedings of Saint Petersburg Electrotechnical University "LETI". Issue 5 (2010), 67-72
9. Deripaska, A.O., Zhukova, N.A., Pan'kin, A.V.: Adaptive Selection of Processes of Handling and Analysis Multivariate Measurements in Intelligent Information Systems. In: Proceedings of 13th Russian conference on Artificial Intelligence with international participation, October 16 – 20, 2012, Russia, Belgorod (CAI-2012) (in Russian)
10. Pankin, A.V., Kuzeny, V.V.: Data Harmonization in CIS. In: Proceedings of International Conference of Information Fusion and Geographic Information Systems 2009, St. Petersburg, pp. 63-76, Springer, Berlin
11. Popovich, V.V., Potapichev, S.N., Sorokin, R.P., Pankin, A.V.: Intelligent GIS for Monitoring Systems Development. In: Proceedings of CORP2005, February 22-25, 2005, University of Technology Vienna
12. Zhukova, N.A., Pankin, A.V.: Principles of managing the processing and analysis of multidimensional measurements in IGIS. In: Proceedings of the Information technologies in management, St. Petersburg, October 9 – 11 (2012)
13. Smith, H., Fingar, P.: Business Process Management (BPM): The Third Wave, Meghan Kiffer Press (2003)
14. Popovich, V., Pankin, A., Galiano, F., Potapichev, S., Zhukova, N.: Service-Oriented Architecture of Intelligent GIS. In: SOMAP 2012
15. Korablev, A. A., Pnyushkov, A.V., Smirnov, A.V.: Creation of an oceanographic database for climate monitoring in the North European basin of the Arctic. In: Trudy AANII. Issue 447 (2007). 85-108
16. Ashik, I.: Recent Russian Marine Research Activities in the Arctic Ocean. Arctic Science, International Law and Climate Change. Volume 235 (2012), 59-66

Author Index