ITHEA

International Journal

INFORMATION CONTENT & PROCESSING

2014 Volume 1 Number 4

International Journal
INFORMATION CONTENT & PROCESSING
Volume 1 / 2014, Number 4

IJ ICP is official publisher of the scientific papers of the members of the ITHEA® International Scientific Society

IJ ICP rules for preparing the manuscripts are compulsory.
The **rules for the papers** for ITHEA International Journals as well as the **subscription fees** are given on www.ithea.org .
The papers should be submitted by ITHEA® Submission system http://ij.ithea.org .
Responsibility for papers published in IJ IMA belongs to authors.

International Journal "INFORMATION CONTENT AND PROCESSING" Volume 1, Number 4, 2014

Edited by the **Institute of Information Theories and Applications FOI ITHEA**, Bulgaria, in collaboration with
Institute of Mathematics and Informatics, BAS, Bulgaria,
V.M.Glushkov Institute of Cybernetics of NAS, Ukraine,
Universidad Politechnika de Madrid, Spain,
Hasselt University, Belgium
Institute of Informatics Problems of the RAS, Russia,
St. Petersburg Institute of Informatics, RAS, Russia
Institute for Informatics and Automation Problems, NAS of the Republic of Armenia.

Publisher: **ITHEA®**
Sofia, 1000, P.O.B. 775, Bulgaria. www.ithea.org, e-mail: info@foibg.com
Technical editor: Ina Markova

Printed in Bulgaria

ISSN 2367-5128 (printed) **ISSN 2367-5152 (online)**

MULTI-LAYER KNOWLEDGE REPRESENTATION

Krassimira Ivanova

Abstract: *An approach for knowledge representation based on post-relation type of information bases is outlined in the paper. Explanation starts with remembering the idea of Natural Language Addressing. After that, the idea of Multi-layer Knowledge Representation by Means of Natural Language Addressing is presented.*

***Keywords**: Multi-layer Knowledge Representation; Natural Language Addressing*

***ACM Keywords**: I.2.4 Knowledge Representation Formalisms and Methods*

Introduction

There are a lot of approaches for knowledge representation [Sowa, 2000]. In this work we show an example of knowledge representation by means of Natural Language Addressing (NLA) [Ivanova et al, 2012a; 2012b; Ivanova et al, 2013a; 2013b; 2013c; 2013d; 2013e; Ivanova, 2013; Ivanova, 2014a]. NLA is based on the Multi-domain Information Model [Markov, 1984; Markov, 2004; Markov, 2004a].

To remember the idea of knowledge representation let start from a letter to Philip Jourdain written in 1914 by Gottlob Frege [Frege, 1980]:

> Let us suppose an explorer travelling in an unexplored country sees a high snow-capped mountain on the northern horizon.
>
> By making inquiries among the natives he learns that its name is 'Aphla'. By sighting it from different points he determines its position as exactly as possible, enters it in a map, and writes in his diary: 'Aphla is at least 5000 meters high'.
>
> Another explorer sees a snow-capped mountain on the southern horizon and learns that it is called Ateb. He enters it in his map under this name.
>
> Later comparison shows that both explorers saw the same mountain. Now the content of the proposition 'Ateb is Aphla' is far from being a mere consequence of the principle of identity, but contains a valuable piece of geographical knowledge. What is stated in the proposition 'Ateb is Aphla' is certainly not the same thing as the content of the proposition 'Ateb is Ateb'.
>
> Now if what corresponded to the name 'Aphla' as part of the thought was the reference of the name and hence the mountain itself, then this would be the same in both thoughts. The thought expressed in the proposition 'Ateb is Aphla' would have to coincide with the one in 'Ateb is Ateb', which is far from being the case. What corresponds to the name 'Ateb' as part of the thought must therefore be different from what corresponds to the name 'Aphla' as part of the thought. This cannot therefore be the reference which is the

same for both names, but must be something which is different in the two cases, and I say accordingly that the sense of the name 'Ateb' is different from the sense of the name 'Aphla'.

Accordingly, the sense of the proposition 'Ateb is at least 5000 meters high' is also different from the sense of the proposition 'Aphla is at least 5000 meters high'. Someone who takes the latter to be true need not therefore take the former to be true. An object can be determined in different ways, and every one of these ways of determining it can give rise to a special name, and these different names then have different senses; for it is not self-evident that it is the same object which is being determined in different ways.

We find this in astronomy in the case of planetoids and comets. Now if the sense of a name was something subjective, then the sense of the proposition in which the name occurs, and hence the thought, would also be something subjective, and the thought one man connects with this proposition would be different from the thought another man connects with it; a common store of thoughts, a common science would be impossible.

It would be impossible for something one man said to contradict what another man said, because the two would not express the same thought at all, but each his owns.

For these reasons I believe that the sense of a name is not something subjective (crossed out: in one's mental life), that it does not therefore belong to psychology, and that it is indispensable [Frege, 1980].

The important knowledge in this example is [Ivanova et al, 2013c]:

- The names Ateb and Aphla refer different parts of the same natural object (mountain, let call it *Pirrin*);
- The position of the referred object (mountain) is fixed by any artificial system (geographical co-ordinates, address) which is another name of the same object;
- The names and the address correspond one to another and both to the real object but without the explorer's map, respectively – the explorer's diary, it is impossible to restore the correspondence;
- At the end, the names Ateb and Aphla are connected hierarchically to the name Pirrin and the relations are:
  - Aphla is_a_South_Side_of Pirrin;
  - Ateb is_a_North_Side_of Pirrin.

The knowledge above is unstructured. For automated processing it has to be structured following some information (data) model. Knowledge representation is closely connected to data models, i.e. the information structures used for organizing the information in the internal or external computer memory. In other words, knowledge representation is depended on the storing patterns and program tools for accessing data. Below we will outline an approach for knowledge representation based on post-relation type of information bases starting with remembering the idea of Natural Language Addressing. After that, the idea of Multi-layer Knowledge Representation by Means of Natural Language Addressing will be presented.

Natural Language Addressing

In this research we follow the proposition of Kr. Markov to use the computer encoding of name's (concept's) letters as logical address of connected to it information stored in a multi-dimensional numbered information spaces [Markov, 1984; Markov, 2004; Markov, 2004a]. This way no indexes are needed and high speed direct access to the text elements is available. It is similar to the natural order addressing in a dictionary where no explicit index is used but the concept by itself locates the definition. For this case we use the term: "Natural Language Addressing" (NL-addressing) [Ivanova et al, 2013a].

The idea of NL-addressing is to use encoding of the name both as relative address and as route in a multi-dimensional information space and this way to speed the access to stored information. For instance, let have the next definition: "**Pirrin:** A mountain in the unexplored country with co-ordinates (x, y)".

In the computer memory, for example, it may be stored in a file at relative address "50067328" and the index couple is: ("Pirrin", "50067328"). At the memory address "50067328" the main text, "A mountain ... (x,y)" will be stored. To read/write the main text, firstly we need to find name "Pirrin" in the index and after that to access memory address "50067328" to read/write the definition.

If we assume that name "Pirrin" in the computer memory is encoded by six numbers (letter codes), for instance by using ASCII encoding system Pirrin is encoded as (80, 105, 114, 114, 105, 110), than we may use these codes for direct address to memory, i.e. ("Pirrin", "80, 105, 114, 114, 105, 110").

Above we have written two times the same name as letters and codes. Because of this we may omit this couple and index, and read/write directly to the address "80, 105, 114, 114, 105, 110".

For human this address will be shown as "Pirrin", but for the computer it will be "80, 105, 114, 114, 105, 110".

Till now, NL-addressing has been presented in several publications [Ivanova et al, 2012a; 2012b; Ivanova et al, 2013a; 2013b; 2013c; 2013d; 2013e; Ivanova, 2013; Ivanova, 2014a].

Examples of models for knowledge representation

Maybe the simplest model for knowledge representation is one used for dictionaries and vocabularies.

From our example we may create a simple vocabulary (Table 1):

Table 1. A simple vocabulary

name	*definition*
Pirrin	A mountain
Aphla	The South Side of Pirrin mountain
Ateb	The North Side of Pirrin mountain
(x, y)	Co-ordinates of Pirrin mountain

The vocabulary in Table 1 is good for using by humans but it is not appropriate for processing in the computer. The same knowledge may be represented in the most popular data format – the relational one. The information from Table 1 may be represented by a relation (Table 2).

Table 2. Relational representation of the simple ontology

object	*South_Side*	*North_Side*	Co-ordinates
Pirrin	**Aphla**	**Ateb**	**(x, y)**

Knowledge from Table 1 or Table 2 may be used for creating a simple ontology with four concepts which may be represented by an ontology graph (Figure 1):

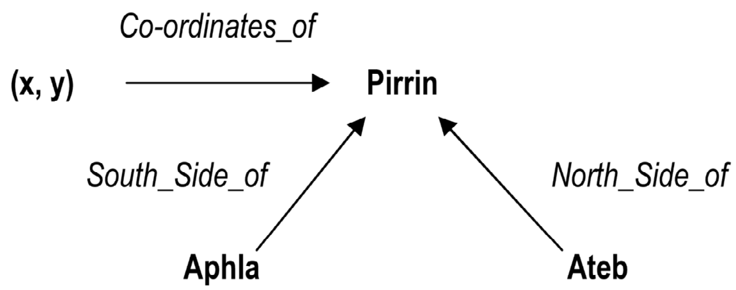


Figure 1. A simple ontology graph

Now we are ready to remember the Resource Description Framework (RDF). It is the W3C recommendation for semantic annotations in the Semantic Web. RDF is a standard syntax for Semantic Web annotations and languages [Klyne & Carroll, 2004]. The underlying structure of any expression in RDF is a collection of triples, each consisting of a subject, a predicate and an object. A set of such triples is called an ***RDF graph***. This can be illustrated by a node and directed-arc diagram, in which each triple is represented as a *node-arc-node* link (hence the term "graph") (Figure 2).

Figure 2. RDF triple

Each triple represents a statement of a relationship between the things denoted by the nodes that it links. Each triple has three parts: (1) *subject*, (2) *object*, and (3) a *predicate* (also called a property) that denotes a relationship. The direction of the arc is significant: it always points toward the object. The nodes of an RDF graph are its subjects and objects. The assertion of an RDF triple says that some relationship, indicated by the predicate, holds between the things denoted by subject and object of the triple. The assertion of an RDF graph amounts to asserting all the triples in it, so the meaning of an RDF graph is the conjunction (logical AND) of the statements corresponding to all the triples it contains. A formal account of the meaning of RDF graphs is given in [Hayes, 2004].

RDF representation of our simple ontology from Figure 1 is given in Table 3.

Table 3. RDF representation of the simple ontology

subject	relation	object
Pirrin	*has_South_Side*	**Aphla**
Pirrin	*has_North_Side*	**Ateb**
Pirrin	has_Co-ordinates	**(x, y)**

From examples given above we may conclude that vocabularies, taxonomies, thesauruses, relations, ontologies, and RDF-graphs, **all what they have in common** are [Pidcock & Uschold, 2012]:

- They are approaches to help structure, classify, model, and/or represent the concepts and relationships pertaining to some subject matter of interest to some community;
- They are intended to enable a community to come to agreement and to commit to use the same terms in the same way;
- There is a set of terms that some community agrees to use to refer to these concepts and relationships;
- The meaning of the terms is specified in some way and to some degree;
- They are fuzzy, ill-defined notions used in many different ways by different individuals and communities.

The **major differences** that distinguish these approaches [Pidcock & Uschold, 2012]:

- How much meaning is specified for each term?
- What notation or language is used to specify the meaning?
- What is the thing for? Taxonomies, thesauruses, and ontologies have different but overlapping uses.

At the end, some additional information may be connected to the names. For instance, it may be the type of mountain, minerals found, some photos, textual descriptions, etc. All such information is connected to names and has to be accessed by names as keywords or paths to it, i.e. its computer representation has to be organized using corresponded pointers, indexes of keyword, etc.

In this case the concept “knowledge representation” is used. As we have seen above, the ontologies are useful approach for knowledge representation, which is understandable for humans as well as for the specialized software.

Example of Multi-layer Knowledge Representation by means of Natural Language Addressing

In this point, we will illustrate an approach for storing RDF-graphs by means of the Natural Language Addressing. Taking in account the interrelations between nodes and edges on Figure 1, a special two-dimensional "multi-layer" representation of knowledge from Table 3 becomes possible (Table 4). It is usual for humans but it is not wide used in the computers.

Table 4. Multi-layer representation of the simple ontology

layer \ object	Pirrin
South_Side	**Aphla**
North_Side	**Ateb**
Co-ordinates	**(x, y)**

The layers form Table 4 may be stored in different files. If we will use the possibility for NL-addressing, the names of the columns will define locations in files of layers.

To receive all knowledge for given node, we have to take node (column) name as NL-address, for instance "Pirrin", and read all information stored at location determined by its encoding ("80, 105, 114, 114, 105, 110") as NL-addresses in different layers (rows "*South_Side*", "*North_Side", and "*Co-ordinates").

In this multi-layer knowledge representation we have one very important achievement – only cells from Table 4, given in bold, will be stored in computer memory. All other information is "virtual" address information used for access to real information. This causes avoiding of supporting indexes for speeding of information search and as a result reducing of used computer resources - memory and processing time.

Note that in practical implementations the tables similar to Table 4 may have several hundred or thousand columns and rows. In addition they are sparse and many cells are empty. NL-Addressing gives possibility to work effectively with such tables not storing empty cells.

Conclusion

Concluding the examples, let point on advantages and disadvantages of the illustrated above multi-layer knowledge representation by means of the Natural Language Addressing.

The main advantages are:

- Reducing the number of tables, which represent the graph;
- Reducing the number of filled cells.

The main disadvantages are:

- The tables are sparse;
- Avoiding pointers we receive a variety of names, which have different lengths and cause difficulties for the implementations in the data bases where the fixed length is preferable;
- The number of nodes may be very great and this way needs corresponded number of columns in the table (in any cases hundreds or thousands).

The disadvantages may be avoided if we will use the Multi-Domain Information Model (MDIM) [Markov, 2004] and corresponded Multi-Domain Access Method (MDAM) [Markov, 1984]. We upgraded MDAM to NL-addressing approach to apply for storing graphs. The possibility to use coordinates is good for graph models where it is possible to replace search with addressing. Hence, the advantages of MDAM are:

- The possibility to build growing space hierarchies of information elements;
- The great power for building interconnections between information elements stored in the information base;
- The practically unlimited number of dimensions (this is the main advantage of the numbered information spaces for graphs where it is possible "*to address, not to search*");

The NL-addressing and multi-layer organization of the information are good basis for implementing this approach for real solutions.

Bibliography

[Frege, 1980] Gottlob Frege An extract from an undated letter, published in Frege's Philosophical and Mathematical Correspondence (ed.) Gottfried Gabriel, Hans Hermes. Friedrich Kanbartel. Christian Thiel, and Albert Veraart, Abridged for the English (edn.), by Brian MeGuinness, and Trans. Hans Kaal (Oxford: Blackwell. 1980), http://mind.ucsd.edu/syllabi/00-01/phil235/a_readings/frege_jourdain.html (accessed: 15.11.2012).

[Hayes, 2004] Patrick Hayes, Editor, RDF Semantics, W3C Recommendation, 10 February 2004, http://www.w3.org/TR/2004/REC-rdf-mt-20040210/ . Latest version available at http://www.w3.org/TR/rdf-mt/ (accessed: 21.02.2013).

[Ivanova et al, 2012a] Krassimira Ivanova, Vitalii Velychko, Krassimir Markov. “About NL-addressing” (К вопросу о естествено-языконой адрессации) In: V. Velychko et al (ed.), Problems of Computer in Intellectualization. ITHEA® 2012, Kiev, Ukraine - Sofia, Bulgaria, ISBN: 978-954-16-0061 0 (printed), ISBN: 978-954-16-0062-7 (online), pp. 77-83 (in Russian).

[Ivanova et al, 2012b] Krassimira Ivanova, Vitalii Velychko, Krassimir Markov. “Storing RDF Graphs using NL-addressing”, In: G. Setlak, M. Alexandrov, K. Markov (ed.), Artificial Intelligence Methods and Techniques for Business and Engineering Applications. ITHEA® 2012, Rzeszow, Poland; Sofia, Bulgaria, ISBN: 978-954-16-0057-3 (printed), ISBN: 978-954-16-0058-0 (online), pp. 84 – 98.

[Ivanova et al, 2013a] Krassimira B. Ivanova, Koen Vanhoof, Krassimir Markov, Vitalii Velychko, “Introduction to the Natural Language Addressing”, International Journal "Information Technologies & Knowledge" Vol.7, Number 2, 2013, ISSN 1313-0455 (printed), 1313-048X (online), pp. 139–146.

[Ivanova et al, 2013b] Krassimira B. Ivanova, Koen Vanhoof, Krassimir Markov, Vitalii Velychko, “Introduction to Storing Graphs by NL-Addressing”, International Journal “Information Theories and Applications”, Vol. 20, Number 3, 2013, ISSN 1310-0513 (printed), 1313-0463 (online), pp. 263 – 284.

[Ivanova et al, 2013c] Krassimira B. Ivanova, Koen Vanhoof, Krassimir Markov, Vitalii Velychko, “Storing Dictionaries and Thesauruses Using NL-Addressing”, International Journal "Information Models and Analyses" Vol.2, Number 3, 2013, ISSN 1314-6416 (printed), 1314-6432(online), pp. 239 - 251.

[Ivanova et al, 2013d] Krassimira B. Ivanova, Koen Vanhoof, Krassimir Markov, Vitalii Velychko, “The Natural Language Addressing Approach”, International Scientific Conference “Modern Informatics: Problems, Achievements, and Prospects of Development”, devoted to the 90th anniversary of academician V. M. Glushkov. Kiev, Ukraine, 2013, ISBN 978-966-02-6928-6, pp. 214 - 215.

[Ivanova et al, 2013e] Krassimira B. Ivanova, Koen Vanhoof, Krassimir Markov, Vitalii Velychko, “Storing Ontologies by NL-Addressing”, IVth All–Russian Conference “Knowledge-Ontology-Theory” (KONT-13), Novosibirsk, Russia, 2013, ISSN 0568 661X, pp. 175 - 184.

[Ivanova, 2013] Krassimira Ivanova, “Informational and Information models”, In Proceedings of 3rd International conference “Knowledge Management and Competitive Intelligence” in the frame of 17th International Forum of Young Scientists “Radio Electronics and Youth in the XXI Century”, Kharkov National University of Radio Electronics (KNURE), Kharkov, Ukraine, Vol.9, 2013, pp 6-7.

[Ivanova, 2014a] Krasimira Ivanova, “Storing Data using Natural Language Addressing”, PhD Thesis, Hasselt University, Belgium, 2014

[Klyne & Carroll, 2004] Graham Klyne and Jeremy J. Carroll, Editors, Resource Description Framework (RDF): Concepts and Abstract Syntax, W3C Recommendation, 10 February 2004, http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/. Latest version available at http://www.w3.org/TR/rdf-concepts/ (accessed: 21.02.2013).

[Markov, 1984] Krassimir Markov, “A Multi-domain Access Method”, Proceedings of the International Conference on Computer Based Scientific Research, PLOVDIV, 1984, pp. 558 - 563.

[Markov, 2004] Krassimir Markov, “Multi-domain information model”, Int. J. Information Theories and Applications, 11/4, 2004, pp. 303 - 308

[Markov, 2004a] Krassimir Markov, “Co-ordinate based physical organization for computer representation of information spaces”, (Координатно базирана физическа организация за компютърно представяне на информационни пространства) Proceedings of the Second International Conference “Information Research, Applications and Education” i.TECH 2004, Varna, Bulgaria, Sofia, FOI-COMMERCE – 2004, стр. 163 - 172 (in Bulgarian).

[Pidcock & Uschold, 2012] Woody Pidcock, Michael Uschold. “What are the differences between a vocabulary, taxonomy, a thesaurus, ontology, and a meta-model?”, InfoGrid - the Web Graph Database http://infogrid.org/trac/wiki/Reference/PidcockArticle, Retrieved November 18, 2012 (accessed: 26.11.2012).

[Sowa, 2000] John F. Sowa, Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks Cole Publishing Co., Pacific Grove, CA, ©2000. 594 + xiv pages; ISBN 0-534-94965-7

Authors' Information

***Ivanova Krassimira** – University of National and World Economy, Sofia, Bulgaria; Institute of Mathematics and Informatics, BAS, Bulgaria; e-mail: krasy78@mail.bg*

Major Fields of Scientific Research: Software Engineering, Business Informatics, Data Mining, Multidimensional multi-layer data structures in self-structured systems

LANGUAGE IDENTIFICATION AND DATING OF THE GREAT PRESLAV INSCRIPTION BASED ON LETTER FREQUENCY

Jordan Tabov, Tzvetan Pavlov

***Abstract**: The Great inscription from Preslav has been cut into the granite column, which has been found during excavations in the city of Preslav.*

It is considered that it contains words of military nature, and numbers written with Greek letters, and is looked at as a part of inscriptions from northeast Bulgaria, called military inventory inscriptions. In modern science dominates the view that these inscriptions have been left from the 'early Bulgarians', settled in Dobrudzha that went to the south cost of Danube with the ruling dynasty of Isperih.

Through comparison between the letter (sound) frequencies in the text in the Great inscription from Preslav with the frequencies of the letters of the names (personal - of rulers, clans and calendar names) in the Nominalia of the Bulgarian rulers, we can deduct, that it is very likely that the language of the Big inscription from Preslav is different from the language of the people ruled by Isperih. This point is towards the unbiased search for landmarks for attribution and dating of the Preslav inscription in broader chronological terms. Additional considerations give us the basis for a hypothesis, that the Great Preslav Inscription is an artifact created during the XV century.

*K**eywords**: inscription from Preslav, frequencies of the letters, dating*

***ACM Classification Keywords**: I.6 Simulation and Modeling, I.6.3 Applications*

1. Introduction

The Great inscription from Preslav has been cut into a granite column with 2 m height and diameter of 0.39 m, which has been found during the excavations in the city of Preslav; it is kept in the Preslav museum under the inv. number 3212. It is considered that it contains words in military nature, and numbers written with Greek letters (Figure 1 and Figure 2), and is look at as a part of the inscriptions from northeast Bulgaria, called military inventory inscriptions. In modern science dominates the view that these inscriptions have been left from the 'early Bulgarians', settled in Dobrudza that went through the Danube with the ruling dynasty of Isperih.

This kind of origin has been attributed to different inscriptions, found mainly in Dobrudza and lands near it.

One of them is the so called 'archaic inscription' in the "Great Stone Cross", which was considered to be written recently after the acceptance of Christianity during the age of Boris I, in Cyrillic, but in the language of the 'early Bulgarians', which was different than the Slavic language [Dobrev & Dobreva, 2001, 132-139]. However, further careful analysis shows that this inscription is in Wallachian and its most likely dating is XVII century [Tabov & Todorov, 2006].

This example shows us that the inscriptions, connected with the 'early Bulgarians', must be looked at critically, without biases, without preliminary opinions, according to which they are a part of the historic heritage of the 'glorious' Bulgarian past from the period of the 'dark ages' in the European history.

Here we will present the thesis, that the language of the Great inscription from Preslav is different from the language of the 'people of Isperih', and we will look for its most likely dating from this point of view.

2. Text and translation of the inscription from Preslav

The Great inscription from Preslav has been published for the first time from I. Venedikov [Venedikov, 1946] and has been analyzed by number of researchers (amongst them must be mentioned V. Beshevliev [Beshevliev, 1981]), and has been looked at as description of weapons (armor and helmets).

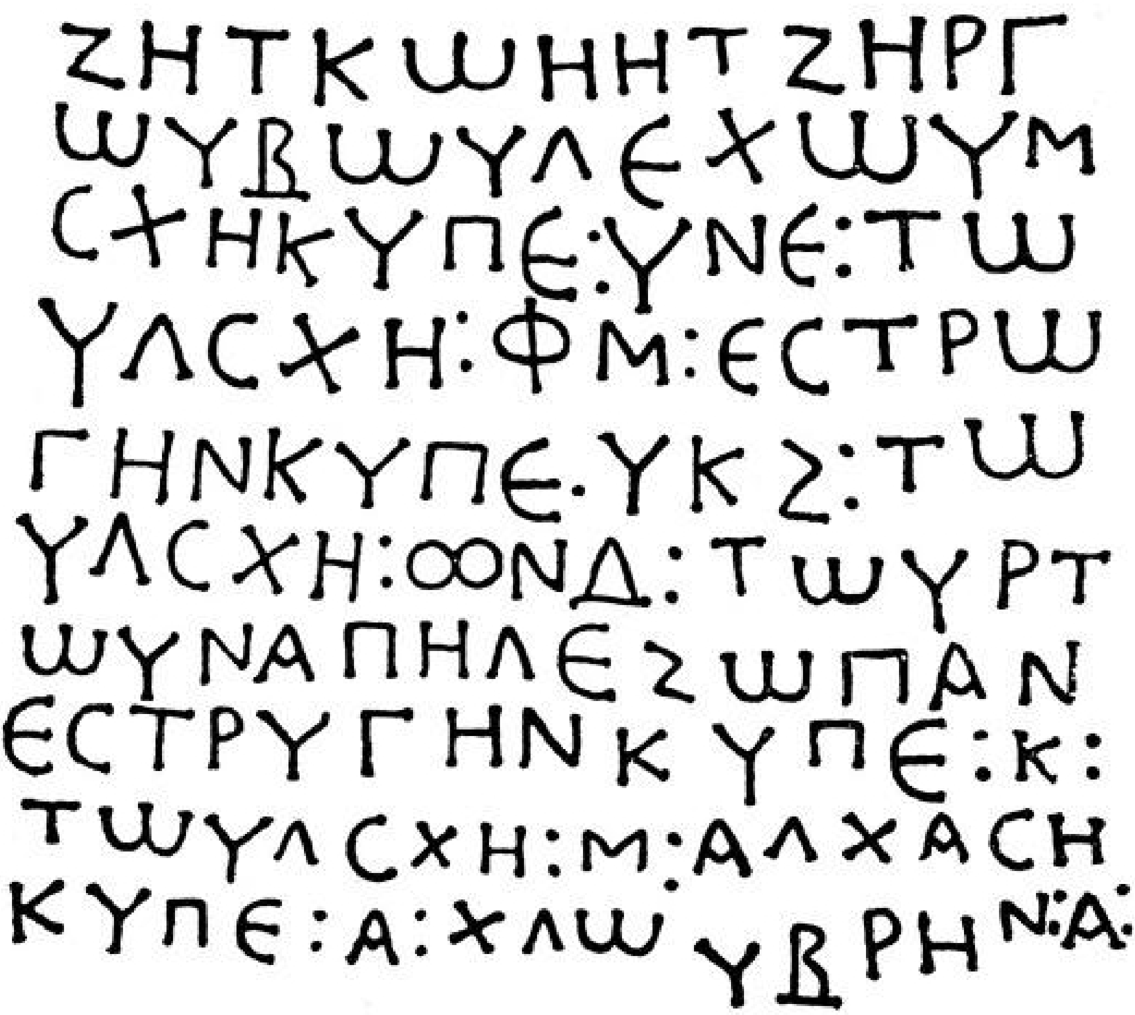


Figure 1. The sketch of the text from the Great inscription from Preslav [Minkova & Ivanov, 2010]

According to [Dobrev & Dobreva, 2001, 88-96] the inscription in Figure 1 - can be sorted as shown in Figure 2 and be translated as below:

> "On the fortified camp ichiguru boila (there are): leather armor 455, helmets 540, knitted armor 427, helmets 854; on the meeting point the zhopan (has): knitted armor 20, helmets 40, armor from rings 1, siege tower 1."
>
> The authors confirm the opinion of Zh. Deni, that the language of the inscription all in all is from Turkic type, and there east Iranian words are predominant [Dobrev & Dobreva, 2001, p.96].

Summing the results from the efforts from the numerous researchers, Minkova and Ivanov [Minkova & Ivanonv, 2010] have come to the conclusion, that “the inscription obviously stays unclear”.

ЖИТКОИ ИЧИРГ

У БУЛЕ ХУМ

ШИ КЮПЕ: ΥΝΕ: ТУ

ЛШИ: ФМ: ЕСТРО

ГИН КЮПЕ: УКZ: ТУ

ЛШИ: &NД: ТУРТ

УНА ПИЛЕ ЖОПАН

ЕСТРЮГИН КЮПЕ: К:

ТУЛШИ: М: АЛХАСИ

КЮПЕ: А: ХЛУБРИН: А:

Figure 2. Transliteration of the text of the Great inscription from Preslav with Cyrillic letters according to [Dobrev & Dobreva, 2001, 89]

We are going to defend the thesis, that the language from the Great inscription from Preslav is different from the language of the “people of Isperih”; to do this we will compare the phonetics of these two languages through the comparison of letter (sound) frequencies.

The method of comparison of the languages from the inscription and Isperih used below partly resembles methods used for determination of the authorship of texts. Methods like these are products of stylometry - the science that treats the task of determination, validation and rejection of the authorship of a certain text.

At first sight it seems that the most natural approach for assessment of authorship differences is the determination of external elements of the style of the given author, mainly his favorite or preferred words or phrases, terms etc. This type of subjective and attributive approach is used to this day. However, the selection of such elements is subjective, and can easily lead to the wrong conclusions, in examples such as deliberate imitation, when outer features of the author have been chosen. Further more in most cases there are missing words and phrases that can be clearly marked as the "author's own". Therefore, it seems that the most fruitful way of exploration of the issues of the authorship of texts is through the search for subconscious features of the language of the author; and they are on the other hand can be detected via suitable formal-quantitative methods [Hmelev, 2014].

3. Statistical methods for the determination of the authorship of a text

While using statistical research with stylometric[1] goals it's assumed that the conclusions of such research can **point or refute authorship only with a certain probability, not complete certainty**; in order to determine the level of probability of the conclusions there is further analysis needed [Buckland, 1999].

One of the pioneers of launching the statistical methods is N. Morozov, who in 1915 suggested [Morozov, 1915] for the statistical regularities of the distribution of specific function words to be researched. The specific parameters of such research (for example, the frequency of usage of the preposition "v") were subject to critics from A. Markov [Markov, 1916], Including the opinion, that in a big volume of text excerpts, all the statistical results (for all authors) will "fluctuate around a middle value, subject to the common laws of the language".

4. Percentage of function words as author's invariant

The Morozov's poroposals have found interesting and important development in the research of V.I. Markova-Fomenko and T.G. Fomenko [Fomenko & Fomenko, 1983], made by A.T. Fomenko's initiative and based upon his specific ideas. They deserve special attention from the view point of our goals.

Based upon A.T. Fomenko's proposal Morozov's idea and Markov's opinion were used for experimental research by V.I. Markova-Fomenko and T.G. Fomenko. For an extensive number of Russian authors they followed the behavior of the following features: 1) the length of the sentences, 2) the length of the words, 3) the frequency of

---

[1] "**Stylometry** is the application of the study of linguistic style, usually to written language, but it has successfully been applied to music and to fine-art paintings as well. Stylometry is often used to attribute authorship to anonymous or disputed documents. It has legal as well as academic and literary applications, ranging from the question of the authorship of Shakespeare's works to forensic linguistics". From Wikipedia, the free encyclopedia.

usage of function words, 4) the frequency of adjectives, 5) the frequency of nouns, 6) the frequency of verbs 7) the frequency of the preposition “v”, 8) the frequency of the particle “ne” (not); the conclusion of this research is, that except the feature 3) all of the above features are not stabilized with the increase of the volume of excerpts, and their values vary in relatively broad limits, therefore the values for a certain author stay comparable (close) to the values of other authors ([Fomenko & Fomenko, 1983; Fomenko, 1980; Postnikov & Fomenko, 1980; Postnikov & Fomenko, 1982]).

The value 3), or namely the percentage of function words in a text, has been named author’s invariant by Fomenko, because this value is constant (with high approximation) in the works of a particular author and, usually, is substantially different for each author. This invariant can be used for the attribution of unknown works and for the detection of plagiarism, however to be used with caution, as wrong deductions are possible, and authors with very similar results for this value can be found (for example Leonov and Fadeev).

Therefore, the value 3) – the frequency of the usage of function words – can be used for the determination of the authorship of a text; we should however clarify – that only within the Russian language. Whether this method can be used for such purposes in other languages it must be researched individually.

5. The problem of the authorship of “And Quiet Flows the Don” (“Tikhiy Don”)

In 1984 there was a book published from several Norwegian and Swedish scholars [Kjetsaa et al, 1984], on the matter of one of the most acute literary question of the 20th century, baring heavy political load; the suspicion of the authorship of one of the brightest works of the Russian literature – the novel “And Quiet Flows the Don”. The book [Kjetsaa et al, 1984] contains studies that confirm the authorship of M. Sholokov and refute the attempts to point a different author of the beginning part of “And Quiet Flows the Don”. The authors explore the distributions of word classes, the usage of combination of grammar classes, the length of sentences, the length of words and others. This way based on many features they confirm the authorship of Sholokhov [Hmelev, 2014].

The approach in the studies [Fomenko & Fomenko, 1983] is substantially different than the approach in the book [Kjetsaa et al, 1984]. The role of the “author’s characteristic” is assigned to the above described “feature 3)”: the percentage of all function words (prepositions, unions and particles) in a coherent fragment of 16 thousand words.

This percentage (“author’s invariant”) is different for each individual author, and its values are between 15% to 30% (here we are looking specifically at Russian authors and works from the Russian literature). This has given the opportunity of the authors of [Fomenko & Fomenko, 1983] to acquire a serious argument for plagiarism from Sholokhov’s side, as the author’s invariant (the value of the “feature 3”) for the first parts of “And Quiet Flows the Don” (books I and II and the beginning of book III), is 19,55 ([Fomenko & Fomenko, 1996, p. 805]), which is significantly different from the author’s invariant in all of the other Sholokhov’s works, including the rest of “And Quiet Flows the Don”, that varies from 22,46 to 24,37 [Fomenko & Fomenko, 1996, p. 805]. This way the study [Fomenko & Fomenko, 1983] justifies the conclusion, that Sholokhov has used an additional source of information for the first parts of the epic.

A further question is addressed in [Fomenko & Fomenko, 1996]: whether this «additional source of information» could be a manuscript of another author from the first half of the 20th century – F. Kryukov. From the stand point of our goals it is an important question, as the volume of Kryukov's works available for analysis was relatively low.

Their author's invariant is 21, 11 [Fomenko & Fomenko, 1996, p. 814] and is different from the author's invariant of the first two books of "And Quiet Flows the Don", however, not as much as the author's invariant of the second part of "And Quiet Flows the Don". The conclusion of [Fomenko & Fomenko, 1996, p. 815] is "M.Sholokhov's invariant is much further from the first two books of the novel, than F. Kryukov", as the authors note, there is a reason for the presumption that Kryukov might be the author, however in order for this to be proven more works need to be analyzed.

6. The Letter Frequencies in Old Bulgarian texts

Basic quantitative characteristics of the letter frequencies in Old Bulgarian texts from different editing's were published for the first time in [Dobreva, 1999]. In this publication they were used for comparison and grouping (through cluster analysis) of manuscripts, that were used for identification and classification of the manuscripts to different literary schools (or traditions) in the past.

One of the goals of this research is to determine the minimum volume of a text excerpt needed for consistent results of the frequency of the usage of letters.

The variations of letter frequencies are different for different letters; in addition they depend on the "volume" of the excerpts (fragments of text, used for the calculation). In [Dobreva, 1999] they are calculated from excerpts of 1 kiloliter (i.e. 1000 letters) – the preferred amount of letters in many of today's philological studies [Dobreva, 1999, 57]. Some of Dobreva's acquired results for medieval oldslavic manuscripts are presented in Table 1 – on the right hand side for Bulgarian manuscripts editing and Russian on the left hand side. In Table 2 we have the data from all the manuscripts from Dobreva's research – Bulgarian, Russian and Serbian editing.

The letters in these tables are sorted according to the size of the standard deviation: the stronger the frequency of a certain letter varies, the higher it is in the table. For example Table 1 – on the left hand side starts with the letter "И", which has an average frequency of 5.44 and standard deviation from this frequency 3.15. The Russian manuscripts studied from Dobreva had shown a higher uniformity: in them the highest standard deviation of the letter frequency is 1.02 (for the big nasal speech sound). It is important to note that the letters **В, Д, М** and **Р**, which frequencies we will be using later, are missing in Table 1. This shows that their frequency is "stable" and varies in small amounts.

For comparing texts (few texts) Dobreva has reached the conclusion, that the more the texts are similar in their origin, the greater the amount of the excerpt needs to be included in the research [Dobreva, 1999, 88-89].

This is because, when similar in origin, close values for the letter frequencies are expected, and with small differences in these values the deviations from the "average amounts" need to be reduced to their minimum.

This logic can be expressed in another way: ***with high differences between the frequencies the texts are not "close according to letter frequencies" and smaller excerpts can be used for groundwork of the conclusions.***

With diverse – "distant"- excerpts the deviations from the average values of letter frequencies are higher: this can be seen in Table 2.

Table 1. Letters with the highest standard deviations in frequent usage – in medieval Old Slavic manuscripts: Bulgarian editing (on the left) and Russian editing (on the right) [Dobreva, 1999, 69 and 68]:

Буква	Средна стойност	Минимална стойност	Максимална стойност	Стандартно отклонение
и	5.44	1.65	9.30	3.15
і	2.51	0.04	6.12	2.40
ѩ	1.24	0.00	2.86	1.24
ѧ	0.88	0.00	1.98	0.86
ь	1.54	0.20	2.73	0.75
ѣ	3.57	2.70	4.80	0.69
ж	1.77	0.80	2.74	0.53
ї	0.50	0.00	1.85	0.51
ꙗ	0.49	0.00	1.21	0.50
о	8.87	7.74	10.22	0.45
а	6.04	5.21	7.35	0.38
ѡ	0.43	0.02	0.97	0.36
ѥ	0.26	0.00	1.14	0.30

Буква	Средна стойност	Минимална стойност	Максимална стойност	Стандартно отклонение
ж	0.78	0.00	2.70	1.02
ъ	3.31	1.82	5.52	0.88
ь	2.58	1.40	3.71	0.69
о	8.93	7.42	10.21	0.56
ꙗ	0.88	0.23	1.95	0.56
ѹ	2.12	1.10	3.14	0.55
ѧ	2.20	1.62	3.18	0.40
ю	0.70	0.07	1.38	0.39
н	8.35	7.41	9.21	0.30
т	5.40	4.16	6.38	0.30
ѣ	2.57	2.06	3.40	0.30

Table 2. Letters with the highest standard deviations with high usage – all of the Old Slavic manuscripts – Bulgarian, Russian and Serbian editing's that were studied by [Dobreva, 1999, 70]:

Буква	Средна стойност	Минимална стойност	Максимална стойност	Стандартно отклонение
ъ	4.28	0.00	11.09	2.71
и	7.46	0.16	10.13	2.38
ь	2.82	0.00	7.53	1.76
і	1.11	0.00	7.17	1.74
ѫ	1.41	0.00	4.00	1.03
ж	1.00	0.00	4.18	1.02
ѩ	0.42	0.00	3.86	0.93
ѥ	0.45	0.00	3.45	0.79
оу	1.91	0.46	4.73	0.70
ѣ	2.94	1.61	5.41	0.68
о	8.87	6.29	12.35	0.67
є	6.03	3.95	9.36	0.57
ꙗ	0.80	0.00	2.33	0.56
д	6.46	4.41	8.65	0.55
ъї	0.86	0.00	2.02	0.55

Буква	Средна стойност	Минимална стойност	Максимална стойност	Стандартно отклонение
ъı	0.38	0.00	1.67	0.53
т	5.37	3.07	7.38	0.52
вь	0.54	0.00	2.60	0.51
м	3.25	0.91	6.34	0.49
р	3.43	1.85	5.40	0.42
с	6.29	3.67	8.18	0.42
в	5.59	3.11	7.21	0.41
г	2.53	1.27	4.27	0.40
ї	0.35	0.00	4.35	0.40
ю	0.58	0.00	1.99	0.38
н	4.60	3.22	7.59	0.35
д	3.35	0.00	4.92	0.34
л	3.07	1.09	4.35	0.33
ѡ	0.66	0.00	1.80	0.33

7. Letter frequencies in the inscription from Preslav

We are using the representation of the text of the Great inscription from Preslav in ***Cyrillic*** letters (Figure 2) according to [Dobrev & Dobreva, 2001, 89]. We eliminate the numbers as we are unaware how they were pronounced.

The rest of the text is:

ЖИТКОИ ИЧИРГУ БУЛЕ ХУМШИ КЮПЕ ТУЛШИ ЕСТРОГИН КЮПЕ ТУЛШИ ТУРТУНА ПИЛЕ ЖОПАН ЕСТРЮГИН КЮПЕ ТУЛШИ АЛХАСИ КЮПЕ ХЛУБРИН

It includes 97 letters in total.

Let us examine the frequencies of the letters: among them we can find four: - В, Д, М and Р:

Letter	**В**	**Д**	**М**	**Р**
Amount in the inscription	0	0	1	5
Percentage in the inscription	0	0	1	5

We are going to treat the names (personal - of the rulers, clans and calendar names) in ***The Nominalia of the Bulgarian khans*** ([Nominalia, Wikipedia]) in an analogical way.

8. The text of the Nominalia of the Bulgarian khans

The ***Nominalia of the Bulgarian khans*** (Bulgarian: *„Именник на българските канове"*), also more known as *„Именник на българските ханове*") is a short chronicle, containing the names and clans of some of the early Bulgarian rulers.

Inside it we can find the dates and length of the periods of their ruling. Interestingly, the title "Khan" has not been mentioned next to any of the names listed. The only mentioned title is „княз", or prince, used next to the name of Iperih ("Исперих") and his five predecessors.

The Nominalia of Bulgarian rulers or "Именник на българските ханове" has been found in 1861 from the Russian scholar Alexander Popov during the study of Russian chronicles. There were three Russian transcripts found: the earliest of them – The Uvarov's, dates to the end of 15th century, the rest two – Pogodinov's and the Moscow one are considered to be from the 16th century. Amongst the three there are few differences in the transcription of the names of the rulers. The text (Table 3) of the work has been included in the book *Hellenic and Roman Chronicle („Елински и римски летописец")*, between *The Forth book of the Kings* and the *Chronicle of Georgi Amartol*, without being separated from them.

Table 3. The text of the Uvarov's transcript ["Nominalia", Wikipedia]

Авитохолъ житъ лѣт. т҃. рѡд ему Дуло. а лѣт ему дилѡмъ твирем. Ирникъ. житъ лѣт. р҃и. рѡд ему Дуло. а лѣт ему дилом тверимь. Гостунъ наместникь сьї два лѣта. рѡд ему. Ерми. а лѣт ему дохсъ. втиремь. Курт: ѯ҃ лѣт дръжа. рѡд ему Дуло. а лѣт ему шегоръ вечемь. Безмеръ г҃. лѣт. а рѡд сему Дуло. а лѣт ему шегоръ вемь. сіі е҃ кнѧз. дръжаше кнѧженїе обону страну Дунаѧ. лѣтъ. ф҃. еі҃. остриженами главами. И потѡм пріиде на страну Дунаѧ. Исперих кнѧз тожде и доселѣ. Есперих кнѧз. ѯ҃а лѣт. рѡд Дуло. а лѣт ему верени алем. Тервен. к҃а. лѣто. рѡд ему Дуло. а лѣт ему текучитем. твирем. к҃и. лѣт. рѡд ему Дуло. а рѡд ему дваншехтем. Севаръ. еі҃. лѣт. рѡд ему Дуло. а лѣт ему тохалтом. Кормисошь. зі҃. лѣт. рѡд ему Вокиль. а лѣт ему шегоръ твиремь. Сїи же княз измѣни рѡд Дулов. рекше Вихтунь. Винех. з҃. лѣт. а рѡд ему Оукиль. а лѣтъ ему имаше Горалемь. Телець. г҃. лѣта. рѡд Оугаинь. а лѣт ему соморъ. алтемь. И сїй иного рад. Оуморъ. м҃. днїи. рѡд ему Оукиль а ему дилѡм тоутѡм.

9. The names of the Nominalia of the Bulgarian khans

We remove the "Slavic" words from the text, and we leave only the names (personal – rulers, clan names, and calendar names):

Авитохол Дуло дилѡм твирем Ирник Дуло дилом тверим Гостун Ерми дохс тирем Курт Дуло шегор вечем Безмер Дуло шегор вем Исперих Есперих Дуло верени але Тервен Дуло текучитем твирем Дуло дваншехтем Севар Дуло тохалтом Кормисош Вокил шегор твирем Дуло Вихтун Винех Укил Горалем Телец Угаин сомор алтем Умор Укил дилѡм тутѡм

The letters here are 269.

10. Letter frequencies in the names of the Nominalia of the Bulgarian khans

After counting the letters В, Д, М and Р the results were the following:

Letter	В	Д	М	Р
Amount in the Nominalia	15	13	20	21
Percentage in the Nominalia (approx.)	6	5	8	8

Now we can compare the frequencies with the frequencies of the same letters = sounds В, Д, М and Р in the Preslav inscription (Table 4**):**

Table 4.

Letter	В	Д	М	Р
Amount in the Nominalia	15	13	20	21
Percentage in the Nominalia (approx.)	6	5	8	8
Amount in the inscription	0	0	1	5
Percentage in the inscription	0	0	1	5

11. Conclusion from the comparison of the letter frequencies in the names of the Nominalia of the Bulgarian khans and the Preslav inscription

As we can see from Table 2, in the old Slavic manuscripts the letters **В, Д, М** and **Р** have relatively low deviations (they are in the second part of the table). This makes them suitable for our goals.

If we take into consideration the following:

1. The relatively high differences between the letter frequency of the letters В, Д, М and Р in the The Nominalia or Bulgarian Rulers and the Preslav inscription (Table 4);
2. The effortless spelling and pronunciation of these letters (and their corresponding sounds) in Old Slavic and Greek letters (the letters both the The Nominalia of Bulgarian rulers and the inscription were written in);
3. The low standard deviations of В, Д, М and Р in old Slavic manuscripts, we can make the following conclusion.

It is more probable that the language of the Great inscription from Preslav is different from the language of the "people of Isperih", than it is probable that it is the same.

This conclusion points us towards the unbiased search for more arguments and landmarks for the dating and attribution of the Preslav inscription in broader chronological terms.

12. YNE = 455 or ONE?

The first from the numbers in the Great Preslav Inscription, written in the form "YNE" (Figure 3), can be transcribed as "une", which is ***one*** in Wallachian. If the inscription is indeed a list of military armor (including weapons), it is natural that it will begin with an important weapon or equipment, that is maybe an only one, for which the whole word is written spontaneously and out of respect, meaning it is a single one.

Figure 3. The number YNE (une, one) in the Preslav inscription

This word points us towards another dating of the inscription, different from the considered one: and namely, the end of 14th and the beginning of 15th century, when on the territory of Dobrudzha other than Bulgarians and Greeks, there were Tartars, Gagauz and Wallachians – all of them under the Ottoman dynasty. In this era it seems natural that some military forces use "eastern languages".

13. "Pile Zhopan" („ПИЛЕ ЖОПАН")

The word Zhupan means administrative ruler of a regional union of the southern Slavs in the past [Online Dictionary, жупан]. It has been widespread and in Wallachia in the XV-XVII century. Ill. 4 represents an excerpt from the text of a Charter of Mikhail voevod, ruler of Wallachia, issued in Targovishte in 1418 [Miletic & Agoura, 1893, 332]. In this excerpt the word Zhupan (жупан) is repeated twice. In addition in the third line we find also the name Pile (Пиле), accompanied by the title ***logothet*** (Figure 4). This makes logical the interpretation of "PILE ZHOPAN" („ПИЛЕ ЖОПАН") as "zhupan Pile". In this text along with Pile are mentioned names like Vojko, Radul, Stancho (2 times), which are common Bulgarian names; but because "pile" is a common Bulgarian word, probably it has been used also as a name.

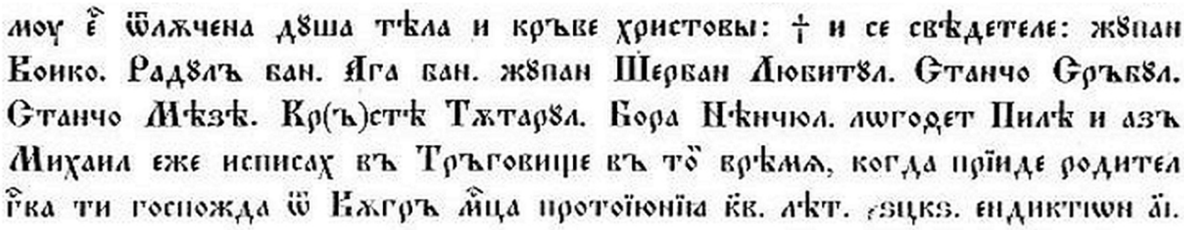

моу є ѿлжчена дѹша тѣла и кръве христовы: † и се свѣдетеле: жѹпан
Боико. Радѹлъ бан. Ѧга бан. жѹпан Шербан Любитѹл. Станчо Сръбѹл.
Станчо Мѣзѣ. Кр(ъ)стѣ Тжтарѹл. Бора Нѣкчюл. лшгодет Пилѣ и азъ
Михаил еже написах въ Тръговище въ то врѣмѧ, когда прїиде родител
гка ти госпожда ѿ Бжгръ мца протоїюніа кв. лѣт. ѕцкѕ. ендиктиѡн аі.

Figure 4. Excerpt from the text of a Charter of Mikhail voevod, ruler of Wallachia, issued in Targovishte in 1418 [Miletic & Agoura, 1893, 332]. The word Zhupan (жупан) is repeated twice

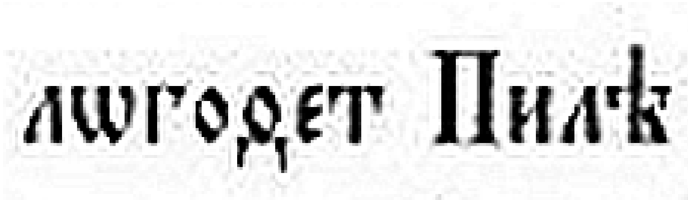


Figure 5. "Logothet Pile" in the text of a Charter of Mikhail voevod (Figure 4)

14. Final argument and our hypothesis

As a final argument for the hypothesis formulated below we can add the following reason: the word "Hlubrin" ("ХЛУБРИН"), which according to [Dobrev & Dobreva, 2001] means "siege tower", according to [Venedikov,1946] – helm, according to [Minkova & Ivanov, 2010] – blacksmith, can be interpreted as Culverin ***"кулверина"*** (in German ***Kolubrine***, in French ***couleuvrine***, in Italian ***colubrine***; from the latin word ***coluber = serpent***, literally means "like serpent") [Culverin, Wikipedia]; this is an early fire arm, predecessor of the musket and arquebus. The earliest report of it dates back to the beginning of 15th century; it is used by the ground troops of France in the middle of 15th century [Culverin, Wikipedia].

The above considerations give the basis for our **hypothesis: The Great Inscription from Preslav is an artifact, created during the 15th century.**

The authors are thankful to Anni Pavlova for her help in the preparation of this article.

Bibliography

[Beshevliev, 1981] Бешевлиев, В. Прабългарски епиграфски паметници. Издателство на Отечествения фронт, София. http://www.promacedonia.org/vb/index.html

[Buckland, 1999] Buckland, W. Forensic Semiotics the Semiotic Review of Books, 1999, 10(3), 10-12

[Culverin, Wikipedia] Culverin. From Wikipedia, the free encyclopedia. http://bg.wikipedia.org/wiki/Кулверина; http://de.wikipedia.org/wiki/Feldschlange; http://fr.wikipedia.org/wiki/Couleuvrine ; http://it.wikipedia.org/wiki/Colubrina

[Dobrev & Dobreva, 2001] Добрев, П., М. Добрева. Древнобългарска епиграфика. Тангра ТанНакРа, София, 2001.

[Dobreva, 1999] Добрева, М. Моделиране на вариативност в старобългарски текстове. Дисертацията за присъждане на научна и образователна степен „Доктор”. Институт по математика и информатика – БАН, София, 1999.

[Fomenko & Fomenko, 1983] Фоменко, В. П., Т. Г. Фоменко. Авторский инвариант русских литературных текстов. В: Методы количественного анализа текстов нарративных источников. АН СССР, Институт Истории СССР, М.осква, 1983. С. 86-109.

[Fomenko & Fomenko, 1996] Фоменко, В. П., Т. Г. Фоменко. Авторский инвариант русских литературных текстов. В: Новая хронология Греции: Античность в средневековье. Т. 2. Изд-во МГУ, Москва, 1996, с. 768 – 820, http://lib.ru/FOMENKOAT/greece.txt

[Fomenko, 1980] Фоменко, А. Т. Некоторые статистические закономерности распределения плотности информации в текстах со шкалой. "Семиотика и информатика", вып. 15, Москва, 1980, стр. 99-124.

[Hmelev, 2014] Хмелев, Д. Краткая история разработки методик определения авторского стиля. http://www.rusf.ru/books/analysis/history.htm (visited on 10.01.2014)

[Kjetsaa et al, 1984] Kjetsaa J., Gustavsson S., Beckman B., Gil S. The Authorship of The Quiet Don. Solum Forlag A/ S: Oslo. Humaities Press: New Jersey, 1984. (Рус. превод: Хьетсо Г., Густавссон С., Бекман Б., Гил С. Кто написал „Тихий Дон”? (Проблема авторства „Тихого Дона”) М., 1989.)

[Markov, 1916] Марков, А. А., Об одном применении статистического метода. Известия АН, сер. 6, т. X, вып. 4, 1916. с. 239.

[Miletic & Agoura, 1893, p. 332] Д-ръ Л. Милетичъи, Д. Д. Агура. Бѣлѣжки отъ едно научно пѫтуванье въ Ромъния. Дако-ромънитѣ и тѣхната славянска писменость. СбНУНК, кн. IX. Министерство на Народното просвещение, София, 1893. http://www.promacedonia.org/lm_da/lm_da_dako-romyni_3a.htm#4f

[Minkova & Ivanov, 2010] Минкова, М. и И. Иванов. Върху значението на някои термини от Преславския военно-инвентарен надпис. (Статията е изнесена на Втората национална конференция по археология, история и културен туризъм „Пътуване към България”, Шумен, 14-16 май 2010). Страница за прабългарите. Език, произход, история и религия в статии, книги и музика. http://protoBulgarians.com

[Morozov, 1915] Морозов, Н.А. Лингвистические спектры: средство для отличения плагиатов от истинных произведений того или иного неизвестного автора. Стилеметрический этюд. Известия отд. русского языка и словесности Имп. Акад. наук, Т. XX, кн. 4, 1915.

[Nominalia, Wikipedia] Именник на българските владетели. От Уикипедия, свободната енциклопедия. http://bg.wikipedia.org/wiki/Именник_на_българските_владетели

[Online Dictionary, жупан] Значение на думата жупан. Онлайн Речник. http://talkoven.onlinerechnik.com/duma/жупан

[Postnikov & Fomenko, 1980] Постников М. М., Фоменко А. Т., Новые методики статистического анализа нарративно-цифрового материала древней истории. (Предварительная публикация). Научный Совет по комплексной проблеме "Кибернетика", Москва, 1980.

[Postnikov & Fomenko, 1982] Постников М. М., А. Т. Фоменко. Новые методики статистического анализа нарративно-цифрового материала древней истории, Уч. зап. Тартуского ун-та. Труды по знаковым системам. XV. Типология культуры, взаимное воздействие культур. - Тарту: Изд-во Тартуского ун-та. - 1982, Вып. 576. с.24-43.

[Tabov & Todorov, 2006] Табов, Й., Н. Тодоров. "Архаичният" надпис на Големия каменен кръст. Palaeobulgarica XXX (2006), 3, 88-99.

[Venedikov, 1946] Венедиков, И. Новооткритият в Преслав първобългарски надпис. ИАИ, XV(1946), 146-160.

Information for the Authors

Jordan Tabov – *Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences, Acad. G.Bonchev Str. Block 8, 1113 Sofia, Bulgaria; e-mail: tabov@math.bas.bg*

Major Fields of Scientific Research: Applications of mathematics and informatics in the humanities, Didactics of mathematics and informatics

Tzvetan Pavlov *– Director, Professional School of Construction, 24 Rayko Daskalov str., Kazanlak, Stara Zagora, Bulgaria; e-mail: cip@abv.bg*

Major Fields of Scientific Research: mathematics, linguistics, didactics of mathematics and linguistics

FACIAL EXPRESSIONS ANALYSIS BASED ON A COMPUTER VISION ALGORITHMS

Iurii Krak, Anton Ternov, Vladislav Kuznetsov

Abstract: *In this paper a variety of computer vision algorithms in a context of their application to the problem of study of facial expressions were analyzed. Performance characteristics of computer vision algorithms for tracking markers which were attached to the face in ideal and real conditions were obtained. A list of algorithms' faults that occurred during experiments was given, and ways to reduce an impact of various factors that influenced the experiments were suggested. The results of experiments were used to model facial expressions on a face of a virtual model of a human head.*

Keywords*: facial expression, computer vision, reference points.*

Introduction and problem statement

In studies of facial expressions, researchers often use tools of verbal description that can contain both text descriptions in a form of predicates and different schematic symbols [Krak_1, 2012]. The results of such studies may be different models which describe variety of human mimic conditions [Hanke, 2002, Sutton, 2002], or a variety of components of such expressions [Miller, 2011].

Analyzing the results of these studies, an amount and a quality of the work and the originality of classification of facial expressions should be noted. However, such research is inherent by a drawback: presence of a human factor that occurs when we create a model of facial expressions. Human intellect, of course, in the analysis of complex objects and in amount of samples of images of objects is more powerful than a computational intelligence, but it is inherent in less accuracy, abstraction and simplification. Also worth noting that verbal data which are obtained from such studies are difficult to analyze and compare.

According to the article [Krak_2, 2012], solution to such problems is the use of computational intelligence in such way that it would be possible to obtain a classification of facial expressions and their components.

The problem statement is to develop and to analyze tools of computational intelligence for the next criteria (requirements) satisfaction:

- High speed of computation;
- High level of automation (low level of human impact during processing time);
- Low level of errors caused by primary data.

Methods used for the analysis of the human facial expressions

Speaking about methods used for the classification of facial expressions and their components, it is important to understand that in a process of solving this problem, we intentionally simplify an object of study. Simplification of the object allows us to get the object model and thus solve the problem using available means of computational intelligence.

A model of object in the solution of this problem is a discrete map (projection) of an object in three-dimensional space by a sequence of maps over time (image stream) in two-dimensional space. Model of the object has a lower space dimension than original one, and becomes discrete by time (not taken into account all states, only states at regular intervals).

For such a model of the object, however, a complement is needed in order to have a possibility to move from a two-dimensional space to the object space. This requires at least another projection of the object to obtain coordinates of an object in three-dimensional space. This can be done in two ways: by using the original image directly and calculating surface of the object based on several projections (solving is based on depth maps), or by going to a mathematical abstraction, by reference points (markers) which are used to calculate position of points on a surface of a face.

Marker is an object with known pre-physical properties (color, shape, size), the form of which is invariant to affine transformations such as rotation and tilt. These criteria are satisfied by spherical object whose color is different from properties of the studied object (a human face). Usage of such abstraction is preferred over direct study of a face surface. Because it can be proven by simple geometric transformations that for each element of the object's model (marker) and for each moment of time based on coordinate increments on projections of two images, real coordinate of a point on a surface of a face can be calculated. Thus, this method allows finding an exact solution for a set of the points, but for all other points, a solution is to be found approximately. A positive aspect of this solving method is lowering computational complexity compared to the solution based only on images.

Considering the solution of this problem by using the abstraction of "tokens" we should find a way to calculate coordinates of a marker in two-dimensional space of frame. The problem of calculating the coordinates can be solved in general by computer vision techniques and algorithms.

Algorithms of computer vision

Algorithms of computer vision is a general name for a class of algorithms used to extract objects from the image, which are set explicitly or implicitly, as well as to get information about these objects. These algorithms include several levels of algorithms that are used at various stages of processing images and primary data [Panin, 2011; Forsyth, 2004; Shapiro, 2006]:

– Image filtering algorithms - extracting specific areas from an image by specific criteria: areas of a certain shade, transitions from one shade to another, etc. Depending on a task, different types of algorithms will be used;

– Algorithms of selecting background and an object on stationary background. These algorithms compare footage sequences and infer the presence and direction of a change. These algorithms include methods of separation of an optical flow and block matching methods of pixels samples on an image with a standard image;

– Algorithms of detection of key features: they select areas of high brightness, corners, lines or circles. Modifications of these algorithms are algorithms for computing descriptors of key features. Based on the information about the image, mathematical operations are performed, which form an object "descriptor". This "descriptor" describes in a special way characteristic areas on an image (areas of singular points);
– Key features comparison algorithms on a video stream: based on similarity criteria, key points are compared on different images. Point's descriptors can be also used instead of the key points;
– Reconstruction algorithms – 3d coordinates are found based on a position of an object on different projections of an object.

Such algorithms can be used for various tasks related to image processing: image compression, pattern recognition and tracking of some objects in the image sequences, etc. A remarkable feature of such problems is the uncertainty associated with the properties of the objects. Objects are rarely flat and can be modified depending on their position in relation to spectator, and information about the properties of an object can be obtained either through training or by sufficiently clear criteria that lay down in the algorithms of comparing key features. Apart from them, there is a class of algorithms that do not belong to the class of computer vision algorithms, but which are used in computer vision problems: - *classification and clustering algorithms*: based on known criteria, obtained from the study of objects or based on information about detected features, separation of features to entire groups (classes) which correspond to specific objects is performed.

A choice of algorithms that suit the best for this task can be done experimentally by verifying a few representatives of the classes of algorithms of computer vision. They must fit to certain criteria to be applicable for solving this task.

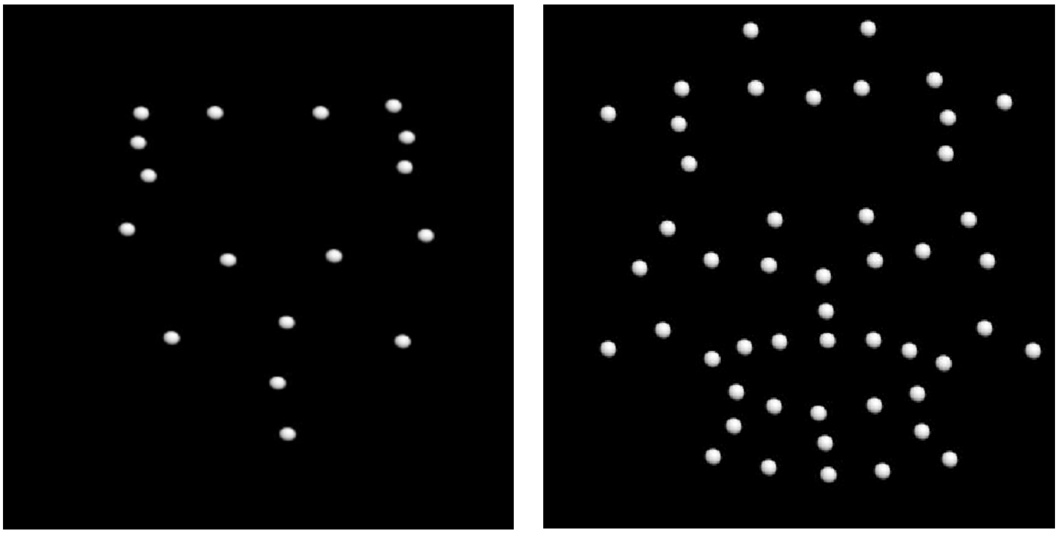
Experimental verification of algorithms. It has been proposed to check computer implementations of these algorithms on real data. The following parameters were analyzed: a) computational speed; b) particular qualities of extraction by different types of algorithms; related to the task of tracking the "marker-like" feature; c) influence of errors in primary data on computation in different algorithms (we did not take in to account errors of primary data during ideal experiment).

Information about the software used in the experiments. Different implementations of computer vision algorithms were tested, in particular: 1) corner detectors Harris, DlineCorner, Susan, Foerstner; 2) key features descriptors: SIFT (Scale-Invariant Feature Transform); 3) optical flow methods: KLT (Kanade-Lukas-Tomasi); 4) block matching methods of pixels' samples in the image with the standard image SAD (Sum of Absolute Differences) and NCC (Normalized Cross Correlation, which contained in the software for video processing: Blender [Blender], ACTS 1.5 [ACTS], Voodoo 1.5 [VisCoda VooDoo], Puchcard VideoTrace [Punchcard VideoTrace], Viscoda VooCAT [VisCoda VooCat], Vicon Boujou [Vicon Boujou], Autodesk Matchmover [Autodesk Matchmover], ImagineerSystems Mocha [ImagineerSystems Mocha], SSontech SynthEyes [SSOntech SynthEyes], Eyeon Fusion [Eyeon Fusion].

Experiments on the study of the algorithms' properties in the ideal conditions

In order to study the performance of algorithms that are present in the software, we created a test video that contained facial expressions without any distortions (including rotation and tilt distortion caused by head movements and some distortions caused by big shutter time in budget camcorders).

A pair of files containing the trajectory of markers on the face was used for these experiments. Files were obtained in professional motion capture recording studio [Easycap facemotion]. First file contained 22 markers (number of markers was reduced in a half), the second one – 48 (original quantity of markers, not taking into account the initial markers on each image). Content of these two files has been converted into a two-dimensional image sequence (video sequence) which represents a frontal view of the three-dimensional scene for each frame appearing on a black background (see Figure 1).



a) b)

Figure 1. Fragments of the first frame of the image sequence: with a reduced number of markers (a); with the original number of markers (b).

Obtained sequences were loaded into a software package in order to get the performance characteristics of individual algorithms used in software implementations. If there was a choice of algorithm (or combination of algorithms) in the software packages, then all the available single algorithms or their combinations were studied (conditional mood).

The experiment was carried out as follows:

- Pre-configured number of tracked features was used – from the minimal 1 (in order to estimate the maximum speed), with several intermediate values of the number of tracked features – 4, 10, 17, 37, 57, 75 (to find out how the estimated performance of the algorithm depends on the number of tracked features) to the maximum value – 75 (total number of features present on the test video sequence);

– Software timer calculated the time required to process the entire video sequence for each run;

– Obtained values were entered into a spreadsheet. Then, the characteristics of the frame rate (FPS), which is used to evaluate the performance of graphic applications, were calculated.

The results of the experiment on the performance at the stage of test run showed that some of the algorithms (or rather their implementations) do not adequately perceive the object "marker". Most of the algorithms which were present in the software packages, such as ACTS, VideoTrace, VooCAT and Voodoo Camera Tracker, detected significantly more (in 3 - 4 times) features that were present in the image. For this reason, we decided to evaluate the performance of the algorithms only in such implementations: 1) in Blender – algorithms: SAD, KLT, Hybrid SAD-KLT, and 2) Autodesk Matchmover – unknown algorithm, and 3) Fusion – unknown algorithm. However, one of the algorithm represented in Autodesk Matchmover showed the best results in quality.

Based on the data obtained in the first experiment, the graphs that show how a processing time depends on the number of features in the image (see Figure 2) were built.

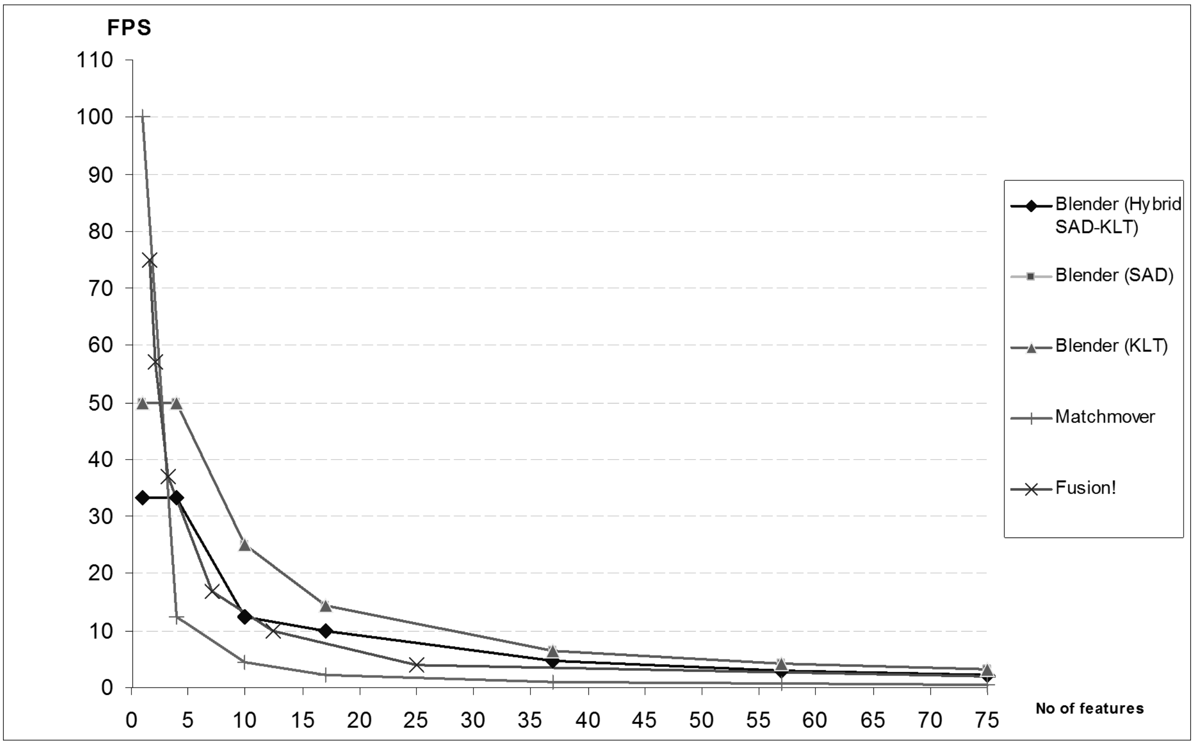


Figure 2. Plots of the processing time depending on the number of features in the image

It can be noted that the implementations of algorithms used in software product Blender showed the best performance characteristics in ideal conditions. However, these graphs do not show the superiority of one algorithm over another, but merely reflect the quality of the algorithms in the specific software package applied for this particular task.

After carrying out the experiment on the calculation of the performance of algorithms' implementations, the errors that might occur during the experiment were analyzed and some suggestions were given:

– Corner detection algorithms (e.g. FAST) find excessive number of points (i.e. circle is perceived as a polygon), and there is no stability (repeatability) in the extraction of features from the same area;
– Feature descriptor algorithms (e.g. SIFT) show better results in conjunction with detectors on which the algorithm is tuned for, otherwise, the algorithms start working with serious errors;
– Algorithms of block matching of pixels samples in the image with the reference image (e.g. NCC) show better results with integral features (such as, e.g., SIFT), than with a set of separate features (e.g., with FAST corner detector);
– Methods of optical flow detection (e.g. KLT) in conjunction with the key features descriptors are faster and more accurate than methods that use a pair of descriptor-descriptor (e.g., SIFT-SIFT).

These findings give a reason to believe that for further experiments on tracking an object "marker" on real data it is desirable to use implementations of algorithms that are better suitable for identifying circular object, such as: 1) circle detection algorithms, 2) algorithms that contain methods of optical flow, and 3) algorithms of block matching of pixels samples in the image with the reference image. This selection was based on a fact that the rest of the algorithms, which were analyzed in this experiment, showed poor results.

Verifying algorithms on real data

For this experiment, a scenario was set up that contained a list of 70 mimic emotional facial expressions and 88 components of these expressions obtained in a study [Krak_2, 2012], and various instructions for carrying out the experiment. Based on this scenario, few videos with different types of markers on the face were taped.

Since to hold such an experiment is very time-consuming (based on the performance of the algorithms presented in Figure 2) it was proposed to conduct the experiment in four stages (Figure 3) for all types of algorithms used in the experiment.

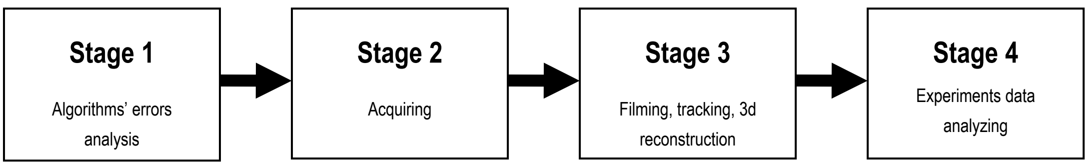


Figure 3. The scheme of the experiment

On the first stage, a study of algorithms' errors of three previously proposed classes was conducted on the real data. Then, a second stage has an objective to analyze the functionality of 3d reconstruction algorithms in order to select the optimal video shooting conditions, which would give adequate results of the 3d reconstruction algorithm. The third stage consisted of: 1) movies shooting based on the shooting conditions obtained on the second stage, 2) analysis of the video clip by computer vision algorithms, which showed the best results on the first stage. The fourth and final stage was an analysis of the experimental results based on the proposed scenario and, thus, evaluations were given to all the experiments conducted in this paper.

Stage 1. Error analysis of algorithms on real data. It was proposed to make a pilot-test on real data of the several algorithms from different classes of computer vision algorithms that are better applicable to the problem of tracking the object "marker":

– The class of optical flow algorithms (for example, the Kanade-Lukas-Tomasi algorithm);

– Class of methods of block matching of pixels samples on an image with the reference image (for example, the Sum of Absolute Differences algorithm);

Class of circle detectors (for example, the algorithm Hough transform, its implementation in the computer vision algorithms library OpenCV [OpenCV]).

For this experiment, few test recordings with a shorter duration were obtained according to the previously proposed scenario. Acquired videos had the following parameters:

– 70 convex markers of the same color, 1000 frames (Figure 4a);

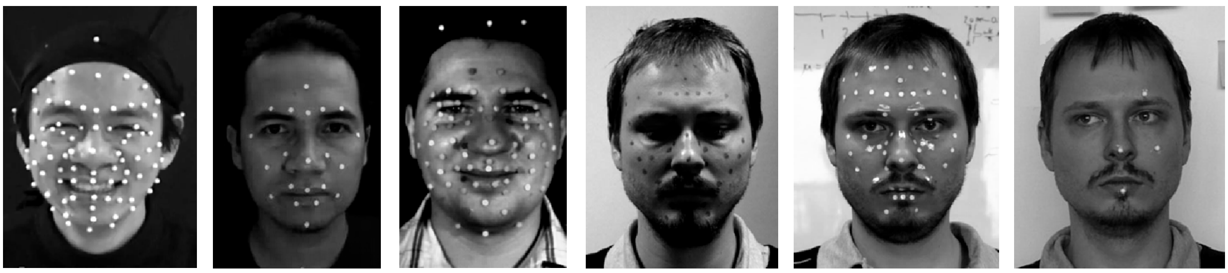
– 20 convex markers of the same color, 750 frames (Figure 4b);

– 70 convex markers of different colors, 350 frames (Figure 4c);

– 22 markers, painted on skin, 1000 frames (Figure 4d);

– 22 flat markers, 1000 frame (Figure 4e);

– 4 convex markers, 1000 frames (Figure 4f).



a) b) c) d) e) f)

Figure 4. Shots from video sequences used in the experiment

During capturing errors analysis, relation between types of used markers and number of primary data errors was analyzed in all cases. Detailed list of the most common errors associated with primary data was compiled.

Errors that accrued due to the type of used marker:

– Usage of flat markers could cause a loss of tracked features if there appear affine transformations such as "turn on the axis Z" and "turn on the axis Y" (so the markers should be round and preferably convex);

- Markers made of washable paint applied to a face might be merged with background color due to camcorder video compression algorithms;
- White spherical markers comparing to flat markers and markers painted by washable paint gave the smallest mean deviation from previous frames;
- The errors occurred when markers appeared on the edge of face and background.

Errors that accrued due to the poor quality of the video:

- Loss or shift of detected feature due to the transition of marker color to the skin color;
- Blurring due to the usage of compression algorithms with a high level of losses during sharp movements;
- Loss of markers may appear on the blurry images of low resolution (less than 640x480p) or because of a hardware compression;
- Auto-Tuning of the brightness and camera focus distance led to the fact that the hue of a feature was also changed and segmentation buy hue value did not work correctly;
- Due to unscattered light from fluorescent lamps, uneven lighting of a face from all sides is observed, which causes a glare on a surface of a face and on markers, which led to errors;
- The presence of noise on an image creates changes of a feature's border location within certain limits (approximately at the diameter of grain of noise spots) that led to the characteristic fluctuations on a mean value.

During the analysis of these algorithms, interesting feature (i.e. ability) in Hough transform algorithm was noted. Even when there is a trail (transition of object "circle" into an object formed by its path) caused by low quality of obtained video, algorithm defined a trail as a single circle, which was located at the "beginning" or at the "end" of the trail.

These results suggest that in addition to the actual raw data error correction (low quality video), the algorithms require completion of noise and color threshold limiter (use of feature's shade instead of brightness of its individual components) or additionally use the method of image segmentation, which would allow to eliminate fluctuations in the brightness of the image. It is also important to note that it's strongly recommended to increase stability of feature's boundary since it leads to a shift of the feature's center location relatively to the real feature location. To enhance the stability of a feature, the adjusting of the results by curve-fitting methods can be used.

Stage 2. Testing the reconstruction algorithms. For testing reconstruction algorithms, a series of experiments were conducted in order to establish the conditions necessary for the shoot. A built-in algorithm to the software Autodesk Matchmover was used for the reconstruction.

Series of experiments were performed with several configurations of cameras

- 1 camera in the middle, one camera on an angle of 45 degrees to the middle camera;
- 2 camera on an angle of 45 degrees to the frontal plane;

– 1 camera in the middle, 1 camera on an angle of 45 degrees lower in the vertical plane.

Results of the experiments have established the following:

– The best visibility of markers was provided when the facial markers were on the background of the skin and the static key markers, which form the global coordinate system, did not stay on the edge of face and background;
– When shooting a face from side angles (1x45 or 2x45) we did not meet the first conditions, so the cameras with such configuration were not used;
– The best result was shown by the configuration of cameras: one camera in the middle, one camera on an angle of 45 degrees lower in the vertical plane.

Stage 3. Movie recording and analysis of the obtained data. Based on the obtained results on a stage 1 and stage 2, the optimal shooting conditions were chosen, however, none of the algorithms fully satisfied all the requirements. Either it had a high accuracy on the ideal data and low error tolerance of primary data, or it had too great drift of a features. In addition, all of them were inherent in the extremely low level of automation - the results have to be converted to other file formats in order to carry out the reconstruction of 3d.

For this purpose, it was proposed to use a software package Autodesk Matchmover, as it fits to the most of the criteria required to obtain three-dimensional coordinates of objects "marker":

As mentioned in the section "**Experiments on study the properties of the algorithms in the ideal conditions**", computer vision algorithms showed the best results of quality (but they seriously lost in speed as shown on Figure 2);

– This software package allows performing both feature tracking and 3d reconstruction of the markers' orientation in space, taking in to account the camera optical distortion and relative position of each camera;
– It was possible to save the data in an XML format that was suitable for the further analysis of the numerical parameters;
– Data obtained from the program can be conveniently used directly in other graphics software packages installed on a computer (for verification of the data).

Besides the Autodesk Matchmover, during the experiments various types of software (including video and audio editing software) were also used.

The process of obtaining of three-dimensional trajectories was held by a particular scenario (see Figure 5).

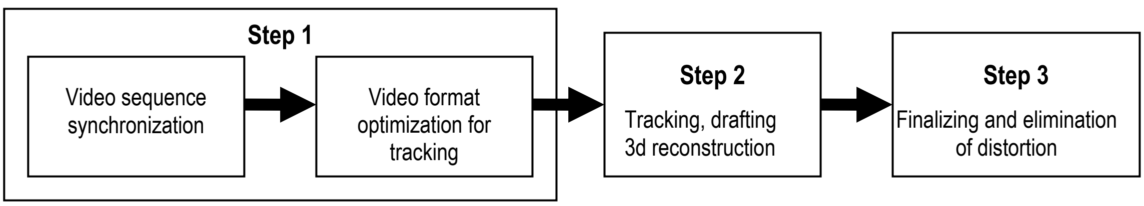


Figure 5. The scenario of a three-dimensional reconstruction

The video obtained from the cameras was separated from audio track. The resulting audio tracks (or rather spectrograms) obtained from each video were analyzed in the program Audacity [Audacity]. Peaks of amplitude (the beginning of a sound) or other features were found on each spectrogram, which proceeded the moment of beginning of useful information on the video. These points on each of the tracks defined the synchronization point of the two videos.

Resulting time values of synchronization points were used during video transformation. Experimentally the best compression algorithm for video was found. It's MJPEG. This format provided algorithms of tracking to work significantly faster comparing to the original video formats MPEG4 or H.264 due to rapid decompression of video in MJPEG (approx. 10 times faster).

Reconstruction is divided into two stages. Initially, the program selects from the entire sequence of images and videos a set of key frames for which the position of each point in space is calculated taking into account a projection of this point on the plane of the frame on each camera. Then obtained results interpolates on the entire range of frames. If some group of markers from the "cloud of markers" changes its position and rotation relatively to the camera, their movements are taken into account in calculating the movements of each marker that belongs to the cloud of markers.

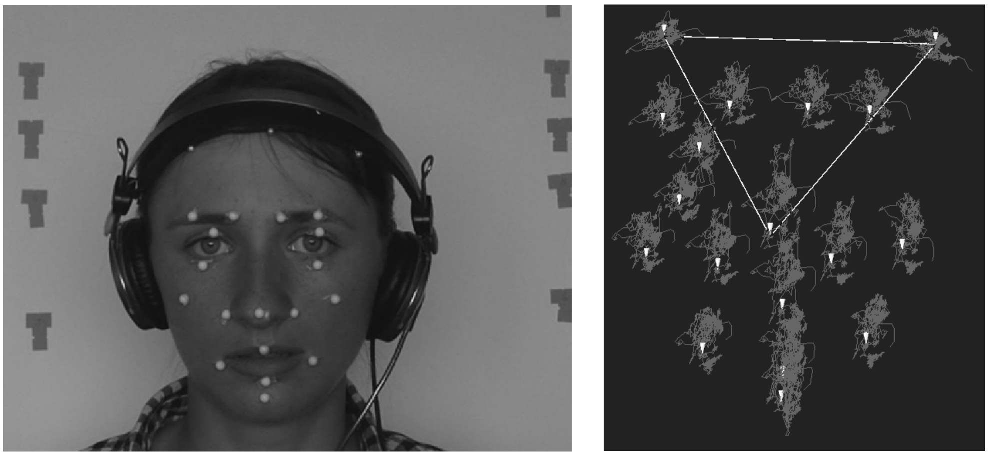
A draft version of the "cloud of markers" has relative position of the markers, which vaguely reminiscent of their location in reality. To obtain more precise location of the markers on a space, calibration of the camera is used (for different versions of the program, this step may be performed before or after the reconstruction of points positions).

Camera calibration is needed for solving such tasks as: 1) clarify the relative positions of cameras in a space, by using the position of static markers in the scene as well (that important), and 2) calculation of the optical distortion of the camera and its settings (even if they are not known in advance). On a stage of calibration, the maximum number of points on a stationary background is calculated and then correction coefficients are calculated for the earlier obtained draft of movements' reconstruction.

On this stage, sufficiently accurate position of each marker (including static) in the space (see Figure 6) is obtained. This can be used for the further studies or to control the facial animation of virtual human head in the one of the computer graphics software packages.

Stage 4. Getting the results of a 3d reconstruction to check the validity of the data. In order to check the quality of 3D reconstruction and to evaluate the reliability of the data, it is necessary to check the data on the test 3d model to view the changes of facial features (jaw, lips, cheeks, nose, eyelids and forehead).

For this purpose, software package Autodesk MotionBuilder [Autodesk Motionbuilder] has been chosen. Obtained data in a cloud of markers were transferred to the resulting parametric face model. Then perception of various facial expressions generated on the face model using the original data was visually rated.

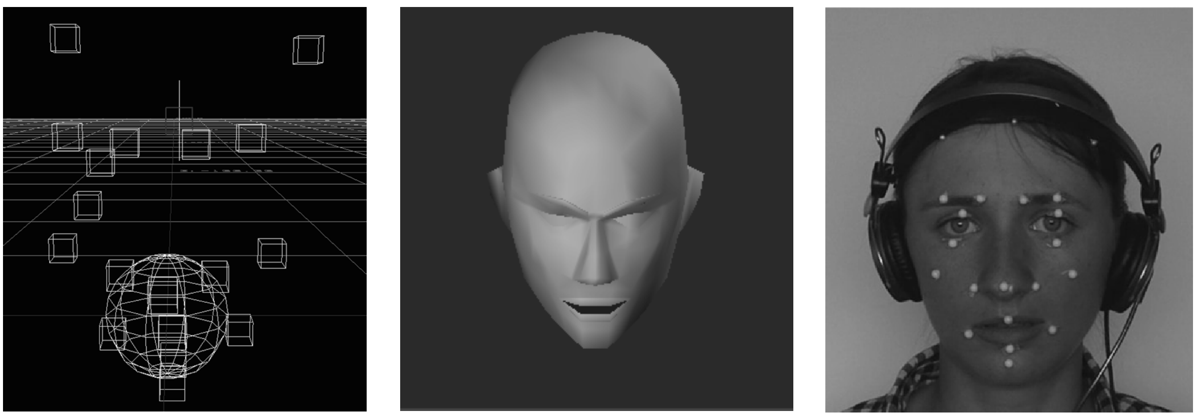


a) b)

Figure 6. Shot from the video sequence (a) and the resulting trajectories of the markers in the software package Autodesk Matchmover (b)

Following steps were made to obtaining synthesized facial expressions. Initial data in the format of three-dimensional trajectories of the points was normalized relatively to the coordinate system. Then, among all the points, several points (three in this case), which did not change the relative position and belonged to a cloud of markers, were found. Based on these points an object “rigid body” was created. The resulting object, particularly its center, was used to create a new reference system (cloud of markers is fixed relatively to the system). The reference system we used to eliminate rotations and tilts of a simplified model of a face. Then, each control point of the simplified model was related to similar point of the marker cloud. These points control the state of facial expressions on a simplified model of a face on each frame of sequence.

These results, when compared with the original video, allowed making a conclusion about a high level of similarity between animated three-dimensional models of facial expressions and facial expressions on the face of a real person captured on a video.



a) b) c)

Figure 7. Comparison of images: cloud of markers a), simplified animated 3d model of a head b) and the original frame of used video c)

Conclusion

The results of this study showed that the specific algorithms should be applied to solve the narrow class of problems. Not all computer vision algorithms are suitable for tracking markers on a human face. Part of algorithms (corner detection algorithms and algorithms based on them) perceives more features than needed and considers them as one. At the same time, a number of algorithms copes with this problem; however, taking in to account the problem of the errors caused by changes of the video settings (brightness variation, noises, etc.), they can be used only with some restrictions.

Computer vision algorithms (including algorithms of 3d reconstruction), used during analysis of the real data and modeling. have shown next features: 1) the reference points method used to study facial expression does allow to calculate accurately changes in some of the points on a face; 2) data obtained by these methods allow to transfer the results to the three dimensional model of a human head and calculate changes for whole points of a face.

Results obtained in this study are considered for further research studies: 1) to simulate facial expressions in sign language by 3d sign language avatar [Krak_3, 2012], and 2) to identify the facial expressions on a known set of facial expressions. It is also proposed to expand the studied set of facial expressions and their components to a larger set in order to cover some of the facial expressions that are specific for Ukrainian sign language.

Bibliography

[ACTS] Web site: http://www.zjucvg.net/acts/acts.html

[Audacity] Web site: http://audacity.sourceforge.net

[Autodesk Matchmover] Web site: http://download.autodesk.com/us/maya/2014help-matchmover/index.html

[Autodesk Motionbuilder] Web site: http://www.autodesk.com/products/motionbuilder/overview

[Blender] Web site: http: //www.blender.org

[Easycap facemotion] Web site: http://www.easycapstudio.com/

[Eyeon Fusion] Web site: http://www.eyeonline.com

[Forsyth, 2004] Forsyth David A., Ponce Jean. Computer Vision: A Modern Approach. – Moscow: Wilyams, 2004 – 928 p. (in Russian)

[Hanke, 2002] Hanke T., Langer G., Metzger C. et al. Interface Definitions, ViSiCAST Deliverable D5-1. – Hamburg, Insitute of German Sign Language and Communication of the Deaf, 2002. – 74 p.

[ImagineerSystems Mocha] Web site: http://www.imagineersystems.com/products/mochapro

[Krak_1, 2012] Krak Iu.V., Ternov A.S., Kuznetsov V.O. System of mimics description for sign language // Artificial Intelligence. – 2012. №4. – P.172-182. (in Ukrainian)

[Krak_2, 2012] Iu. Krak, A. Ternov, V. Kuznetsov, Yu. Barchukova. Creating audiovisual database of nonmanual components for sign language. Eleventh All-Ukrainian International Conference on Signal/Image Processing and Pattern Recognition - UkrObraz'2012. October 15–19, 2012, Kyjiv, Ukraine. Conference Proceeding. – Kyjiv, 2012. – P.7-10. (in Ukrainian)

[Krak_3, 2012] Iu. Krak, M. Lisnyak, A. Ternov. Rozrobka osnovnih instrumentiv kompyuternoi animacii dlya pobudovi sistemi sintezu gestovoi movi. International Conference: Intellectual Systems for Decision Making and Problems of

Computational Intelligence (27–31 may 2012. Kherson, Ukraine). Conference Proceeding. – Knerson:KNTU, 2012. – P.104-105. (in Ukrainian)

[Miller, 2011] Miller F.P., Vandome A.F., McBrewster J. Facial Action Coding System. – Beau Bassin, International Book Marketing Service Ltd, 2011. – 80 p.

[OpenCV] Web site: http://www.opencv.org/

[Panin, 2011] Giorgio Panin and Peter Ratner. Model-based Visual Tracking: the OpenTL Framework. – John Wiley & Sons Inc., New York, 2011 – 318 p.

[Punchcard VideoTrace] Web site: http://punchcard.com.au/

[Shapiro, 2006] Shapiro Linda G., Stockman George C. Computer Vision. – Moscow: BINOM, 2006 – 752 p.(in Russian)

[SSOntech SynthEyes] Web site: http://www.ssontech.com/synovu.htm

[Sutton, 2002] Sutton V. Lessons in SignWriting V. – La Jolla, Center for Sutton Movement Writing, 2002. – 195p.

[Vicon Boujou] Web site: http://www.vicon.com/boujou/

[VisCoda VooCat] Web site: http://www.viscoda.com/

[VisCoda VooDoo] Web site:http://www.viscoda.com/index.php/en/products/non-commercial/voodoo-camera-tracker

Authors' Information

Iurii Krak *– V.M.Glushkov Cybernetics Institute of NASU, senior scientist krak@unicyb.kiev.ua, yuri.krak@gmail.com*

Anton Ternov – *V.M.Glushkov Cybernetics Institute of NASU, scientist, anton.ternov@gmail.com*

Vladislav Kuznetsov – *V.M.Glushkov Cybernetics Institute of NASU, post graduate student, Vlad.Kuznetsov@gmail.com*

address: 40 Glushkov ave., Kiev, Ukraine, 03680

ONTOLOGY OF EDUCATIONAL STANDARDS

Oleksandr Stryzhak

Abstract: *This article discusses the formation of educational standards. The method of forming the content of the educational process describes. The method is based on the ontological knowledge representation and ontological models describing different categories of educational process. Is an example of the formation of educational and vocational programs educational qualification level, the allocation of appropriate skills competencies, creating curricula based on application of an ontological description of subject knowledge.*

Keywords: *ontology, knowledge systems, educational programs, ontograph.*

ACM Classification Keywords*: G.2 Discrete Mathematics – G.2.2 Graph Theory – Trees; I.2 Artificial Intelligence - I.2.4 Knowledge Representation Formalisms and Methods; K.3 Computers And Education – K.3.2 Computer and Information Science Education*

Introduction

Quality of training of skillful specialists in various areas of human activity depends on its content. The content factor of educational process is determined by the following categories:

- National qualification system is an aggregate of mechanisms for legal and institutional regulation of employee's skills taking into consideration the needs of labor market and capabilities of education system;
- Qualification levels are the requirements, determined and adopted under procedure established by legislation, to competence of employee when performing job responsibilities considering their complexity and level of responsibility.

The full-scale content disclosure of such categories is implemented on the basis of national framework of skills (NFS) — systemic and structured description of skills, educational and qualification levels, qualification standards of various levels and types established on the basis of criteria set determined by law. NFS determines the single scale of qualification levels of generally professional competences to develop the sectoral framework of skills and professional standards. NFS ensures inter-sectoral comparability of skills and competences; it is a basis for system with regard to confirmation of compliance and awarding the qualification to specialists. The national framework of skills consists of description of general characteristics of professional activity for each qualification level.

Thereby, NFS determines the dynamics of development of education standards as a definite system of norms and requirements, which describe the mandatory minimal content of main educational and professional programmes, maximal scope educational load and requirements to training level of specialists. Educational and professional programme (EPP) is a sectoral regulation, which determines the regulatory period and content of education, regulatory forms of state assessment, sets the requirements to content, scope and level of education and professional training of specialist by the subject area of training. The educational standard is an element of sectoral standards of higher education and it is used for the following activities:

- Develop and adjust the element of sectoral standards of higher education (diagnostic facilities of higher education quality);
- Develop and adjust the elements of higher education standards of higher education establishments (variation parts of education and professional programme for specialists training and diagnostic facilities of higher education quality, curriculum, programmes of subjects and practices);
- Determine the content of training within the system of retraining and improvement of skills of specialists.

Ontological description methods

Dynamics of creation and development of modern technologies require from education system new approaches when developing NFS. The main requirement of such development should be succession of content and keeping an inter-disciplinarity at the level of subject and topical filling of subjects.

One of such technological approaches is to develop an information base that contensively secures the whole educational process of specialists training in higher school on the basis of ontological simulation of information processes.

As we already mentioned, the foundation of information base to support an education process of any university are curricula, educational programmes and library resources. At present practically all these resources are electronically. Practically all manuals, tutorials, monographs and any other information materials also have their own electronic images. Each such electronic image displays a certain scope of thematic knowledge. Such knowledge, presented in the form of information descriptions as natural and language structures [Гладун, 1994], mirror the considerations about certain subject and thematic facts. The facts are linked with each other by certain relationship and could be also characterized by definite features.

In general, all educational materials and, in particular, the bookish ones are described by scientific style, which is focused on transfer of certain information or explanation of any facts from the scientific point of view. Such style uses various terms and professional vocabulary. The style is typical for educational scientific literature. Scientific style is a style of scientific messages. The area of usage of this style is science and scientific magazines, addressees of text messages could be scientists, future specialists, students and any person who is interests in this sort that scientific area; the authors of tests of the said style are scientists and specialists in their area. The purpose of style one can name the description of laws, identification of regularities, description of discoveries, training etc.

The main function of a style is to communicate information as well as to prove its truth. It is characterized by small terms, generally scientific words, an abstract vocabulary, at theta noun prevails; there is a lot of abstract and material nouns. Judgments have a type of specific expressions and determine the sets of actions, which could be applied in the process of resolving the concrete subject and thematic tasks.

Separation of a set of actions on the basis of knowledge system described and presented in book is possible provided that the procedure of structuring is applied to its natural and language text. To this end we will transform somehow the bookish text having presented it not in the usual form of information description that is consecutive and consistent by style but having represented it in an aggregate of specific statements and expressions. The concrete subject statements/expressions, which have the subject focus, could form the passive base of knowledge.

Transformation of passive base of knowledge, which is represented in the form of information descriptions set forth in the book, into the active is possible on the basis of transformation of such descriptions into the certain terminological fields [Коршунова, 2009], where the concrete notions become concepts of subject are, which is described in a book [Гладун, 1994; Палагин, 2005]. The abovementioned concepts are the certain statements, which determine the specific actions and results of such actions. The statements themselves are designed on the basis of usage of semantics of concepts and those relations, which link these concepts by certain sense.

Sets ofstatements/expressions, which are developed on the basis of subject concepts, form a certain category [Rydeheard, 1988], the objects of which have morphisms for each pair of statement-expression. In future we will consider the morphisms, which transform expressions, given in the text of book, into statements being true for the described facts. One should note that it is difficult to separate the notions “statement” and “expression”; they are practically equivalent. Such separation is fairly artificial and it is constructive in the terms of normal system theory [Malishevski, 1998], where “expression” determines a number of source or passive data; however “statements” allow to assigning and determining the specific dynamic activities. It means that we form a set of concrete subject expressions; then we represent such expressions into the form of statements rephrasing them in an affirmative form [Mendelson, 1997]. Such affirmative form will help us to formulate a certain hypothesis, which on the basis of value of concrete facts representing the displays of concrete phenomena, can be confirmed or to be proved as groundless [Kleene, 1952].

Let us formulate the following rule for expressions and statements:

a) Statement/expression is considered as applicable if there are the conditions determining its truth;

b) Statements/expressionsare considered as inapplicable if there are no conditions determining its truth.

Now we can consider the definition of computer ontology stated in the works of N. Guarino and N. Gruber [Guarino, 1994; Gruber, 1993]–where definition of ontology is considered as specification of conceptualization. The ontology includes also the systems of restrictions, which could be overlaid onto relations within the subject matter of knowledge domain and are expressed in the form of a definite set of axioms that is determined on the basis of notions-concepts and relations between them. Thereby we are able to consider the ontology as conceptually definite and established system of knowledge.

Constructively, ontology as a definite category of development and implementation of information technologies we will understand and form on the basis of definition given in the said work [Гладун, 1994; Палагин, 2005]:

- Hierarchic structure of final aggregate of notions-concepts, which describe the set knowledge domain (KD);

- Structurally ontologycan be represented in the form of ontograph, the tops of which are notions and arches are semantic relations between them;
- Notions-concepts and relationship are interpreted in accordance with the generally valid functions of interpretation taken from electronic sources of knowledge of the set KD;
- Notions and relations are determined on the basis of axioms and restrictions (rules) of their area of action;
- There is a means of formal description of ontograph;
- Functions of interpretation and axioms could be described in notation of formal theory.

Thus, ontology means something more than just detailed set of notions and relations. Therefore, one can describe ontology as an active system of knowledge, which includes an aggregate of objects, and having linked them with descriptions as well as having introduced the formal axioms restricting interpretation and joint usage of such terms. It means that ontology can be considered as some logical theory, some calculation with its own rules. This theory allows to classifying the categories of reality and/or those being expressed in the language of value. [Гладун, 1994].

In general form an ontological model can be formed and represented on the basis of the following four categories:

- Notions-concepts;
- Relations and features;
- Axioms;
- Functions of Interpretation.

Notions are considered as conceptualization of class of all representatives of some essence or phenomenon (for instance, SOFTWARE ENGINEERING, INFORMATION TECHNOLOGY, COMPUTER NETWORKS, DESIGNING OF LOCAL NETWORKS, DATA BASES, and RELATIONAL ALGEBRA etc.). Classes combining notions-concepts on the basis of certain features and relations are general categories, which could be regulated hierarchically. Each class describes the group of individual essences, which are combined on the basis of common features. The most widespread type of relations being used in all ontologies is relationship of categorization, i.e., attribution to certain category. On the basis of this type of relationship the taxonomy of text document is being set [Grossi, 2005]. Axioms and functions of interpretation specify the conditions of attribution of categories and relations and they express the concrete statements.

Now we can consider the process of development of educational and professional programm of educational and qualification level BACHELOR with respect to speciality SOFTWARE ENGINEERING and corresponding qualification SPECIALIST ON SOFTWARE DESIGNING AND TESTING.

Educational Program Ontology - example

The papers [Гладун, 1994; Палагин, 2005; Величко, 2009] describe methodologies and procedures for transformation of passive knowledge system represented in book into active ontology, which ensures information support of subject tasks of teacher and student. Thus, the Figure 1 shows a fragment of ontograph, the tops of which are notions-concepts disclosing the educational and qualification characteristics of speciality SOFTWARE

ENGINEERING. Figure 2 shows a fragment of ontograph, the tops of which determine concepts of main functional areas of the said speciality. Figures 3-a and 3-b represent a fragment of ontology of knowledge system determining the content of the respective subject and topic disciplines being used for training of specialists.

Figures 4a – 4b show the content of a top-concept on discipline SOFTWARE DESIGNING, give examples of inclusion of the respective information content of the course. Figures 5a – 5b represent a fragment of ontology related to development of credit module with regard to discipline SOFTWARE OF DATABASES being one of the courses with regard to discipline SOFTWARE ENGINEERING.

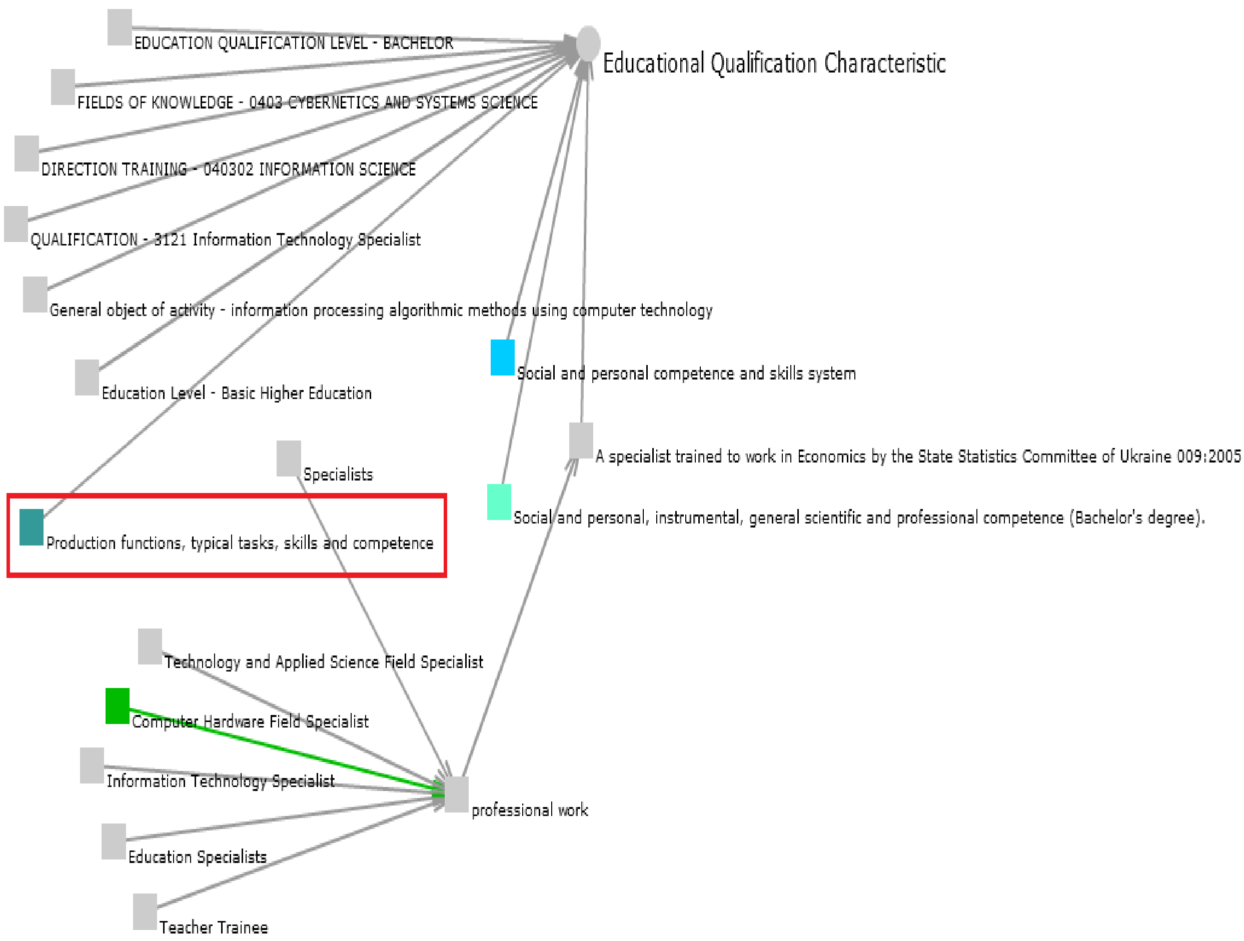


Figure1. Fragment of ontograph of educational and qualification characteristics for speciality SOFTWARE ENGINEERING

As one can see, ontograph means educational and qualification characteristics for speciality, which the skills of specialist should comply with. Ontograph should understand the business functions, be knowledgeable about modern information technologies, have the respective system of skills etc.

Let us consider the top PRODUCTION FUNCTIONS, MODEL TASKS OF ACTIVITY, SKILLS AND COMPETENCES of graduate. The content of this top-concept is shown in Figure 2 in the form of ontograph of main functional areas to be inherent in graduate.

The given fragment of ontology shows functions of software designing. The tops-concepts in this case are main skills of qualified specialist.

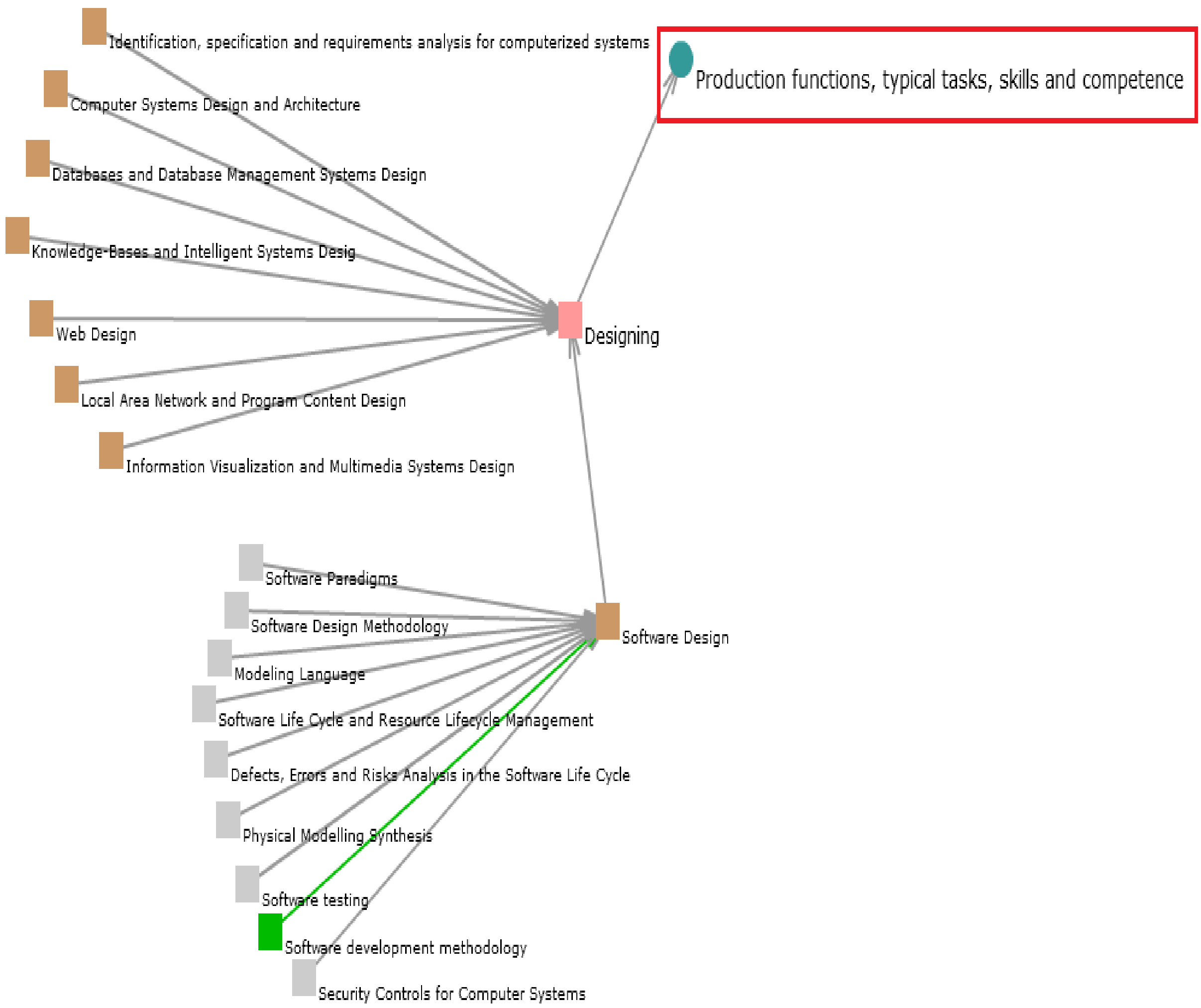


Figure 2. Fragment of ontograph of main functional area of speciality SOFTWARE ENGINEERING

The ontology shows taxonomy dependence between all categories, which determine the level of training in the area of usage of software designing mechanisms.

Figures 3a – 3b show ontology of knowledge systems (passive) on subject and topical disciplines, which comprise the complete course with regard to speciality SOFTWARE ENGINEERING.

Each top-concept determines the content of such disciplines. This content can be represented by a number of manuals, educational and methodological textbooks and monographs on subjects of concrete discipline. An example of substantive filling of ontograph tops is given in Figure 4a – 4b.

Figure 5a – 5b shows organization of credit module on one of each disciplines of speciality. The main types of educational activity on concrete subject disciplines are determined. Each discipline composes the system of knowledge on speciality. Educational load, curricula and educational programmes are being determined. And now each programme and each curriculum are supported by the corresponding information resource – monographs, methodological recommendations, tutorials etc.

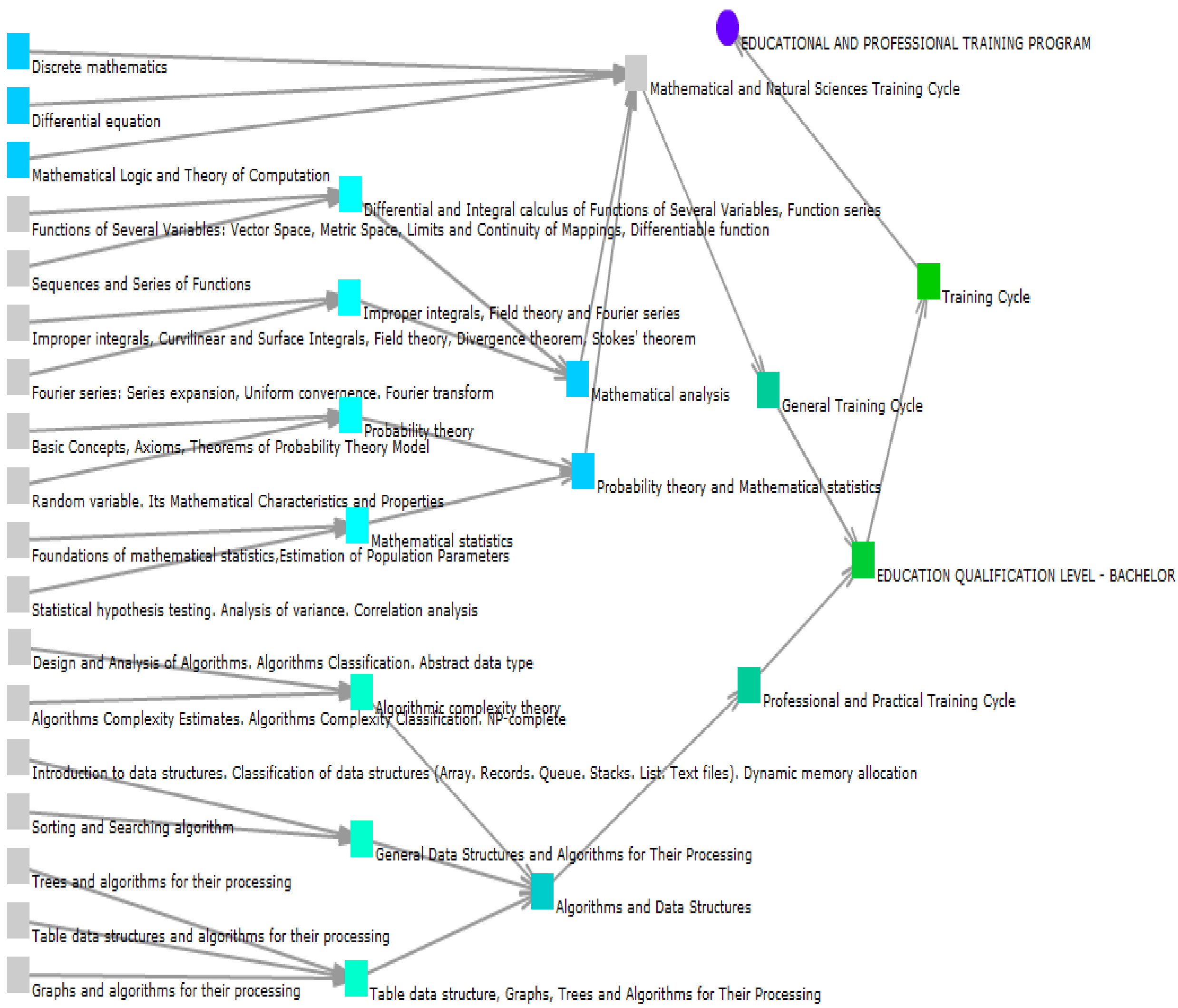


Figure 3a. Fragment of ontology of subject and topical disciplines on speciality SOFTWARE ENGINEERING.

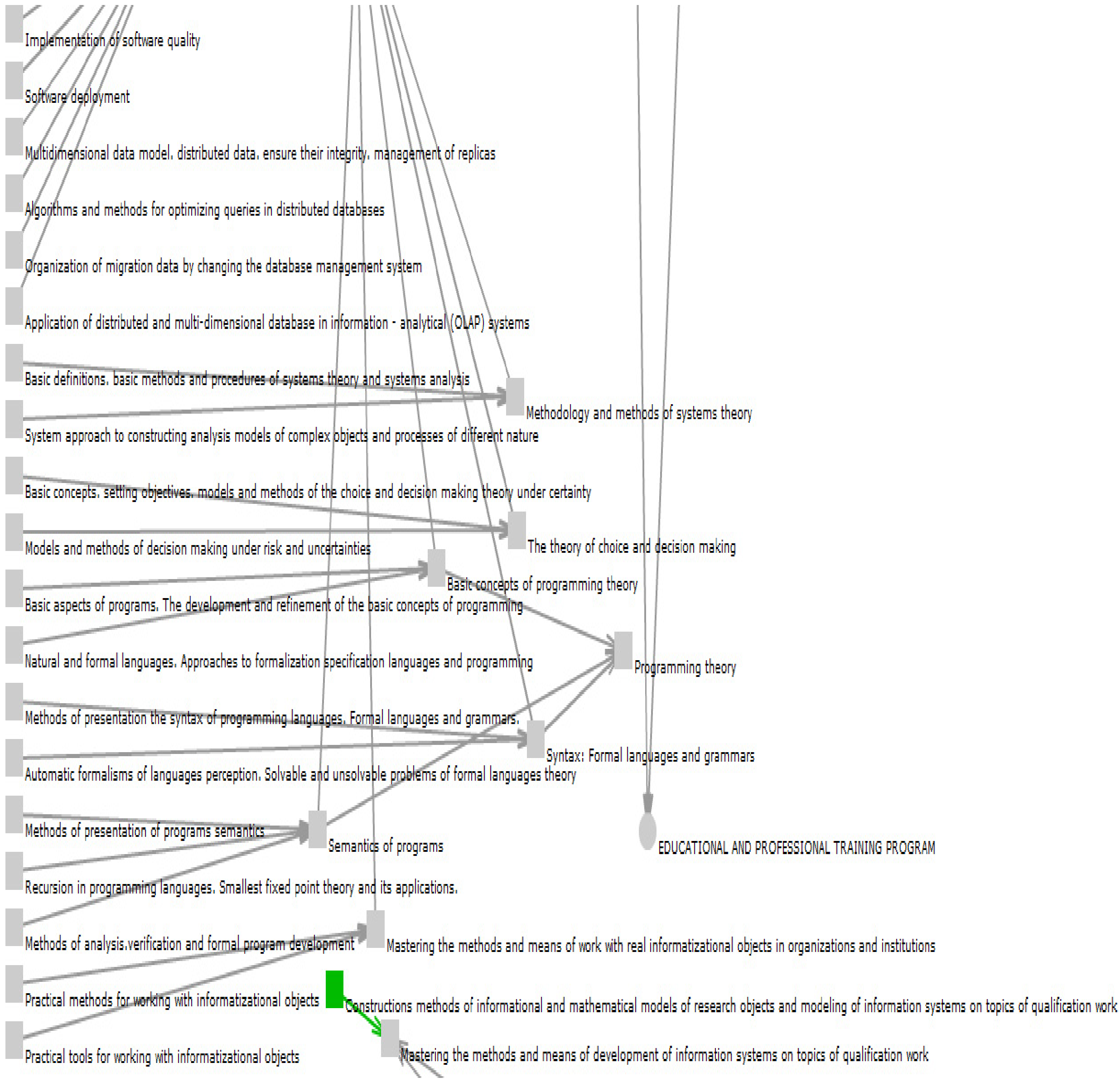


Figure 3b. Fragment of ontology of subject and topical disciplines on speciality SOFTWARE ENGINEERING (continued)

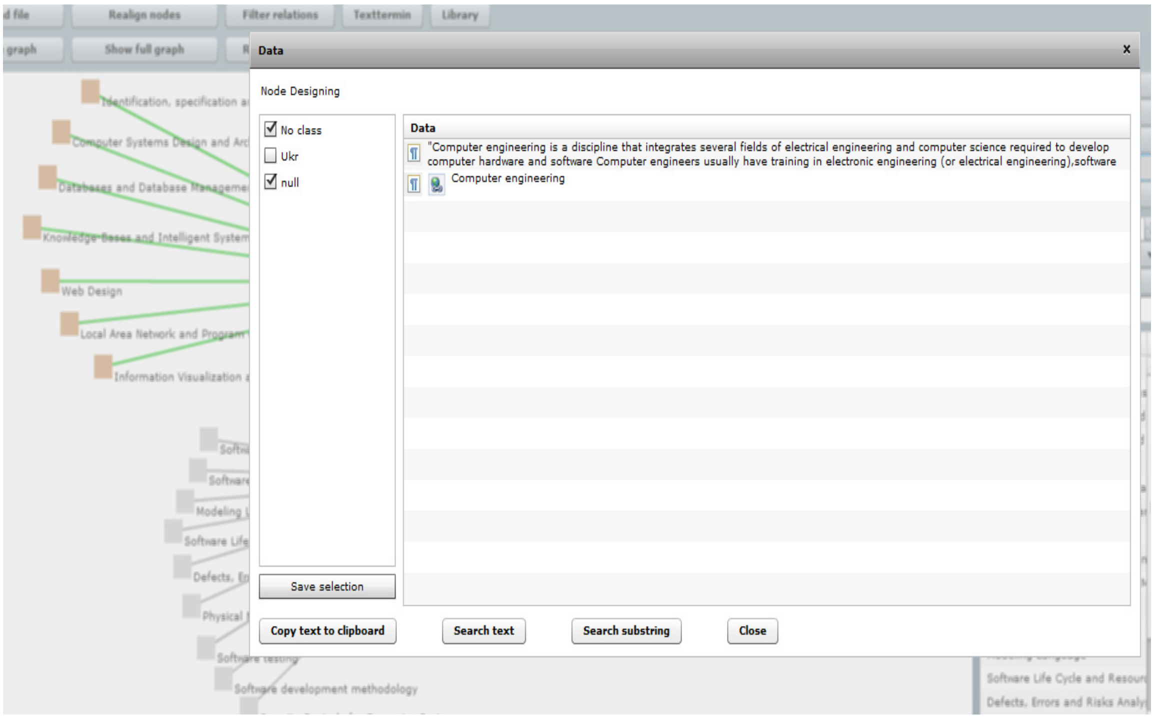


Figure 4a. Content of top-concept on discipline SOFTWARE DESIGNING

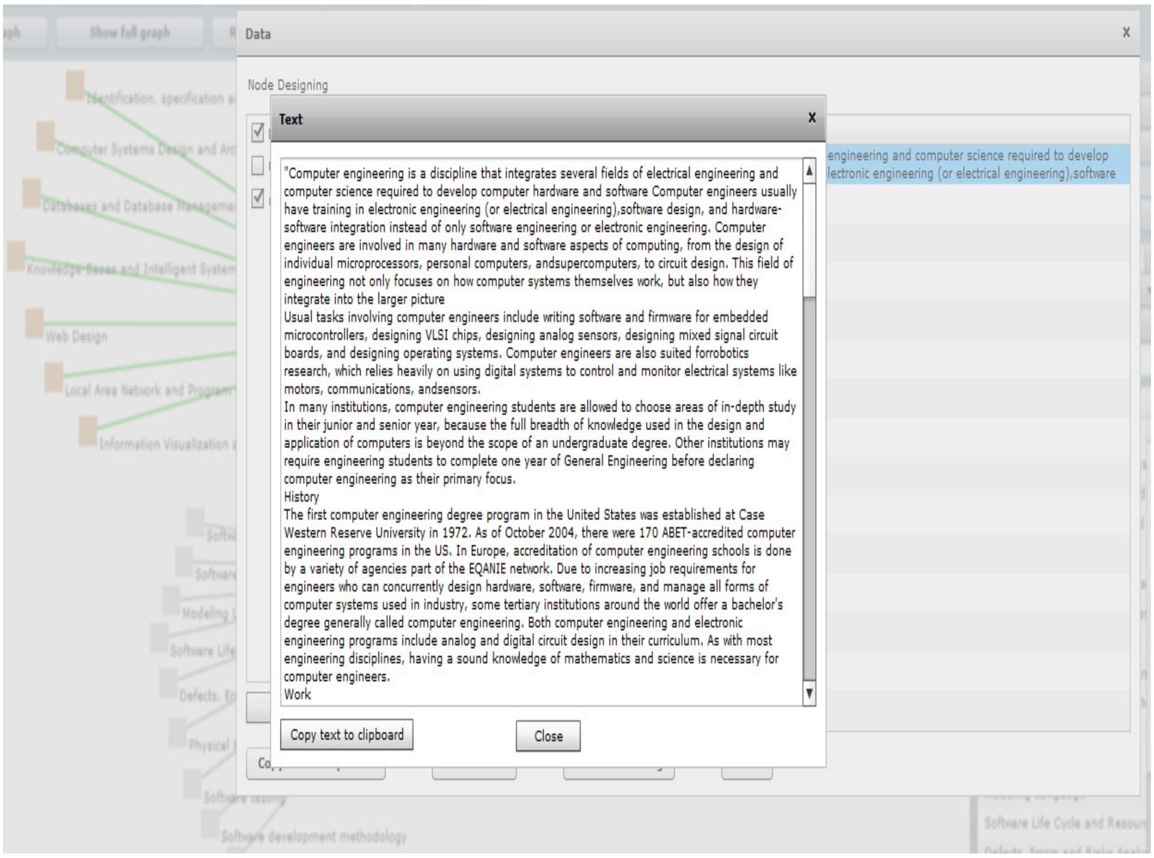


Figure 4b. Description of top-concept on discipline SOFTWARE DESIGNING

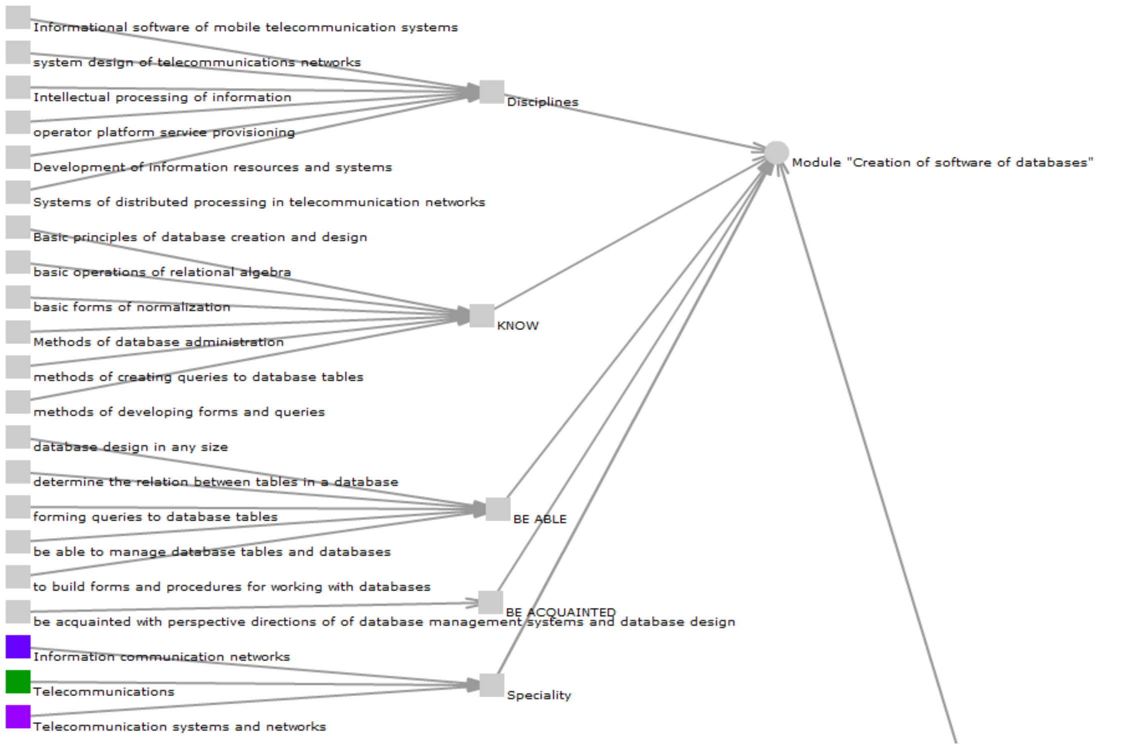


Figure 5a. Fragment of ontology to develop the credit module on discipline SOFTWARE OF DATA BASES

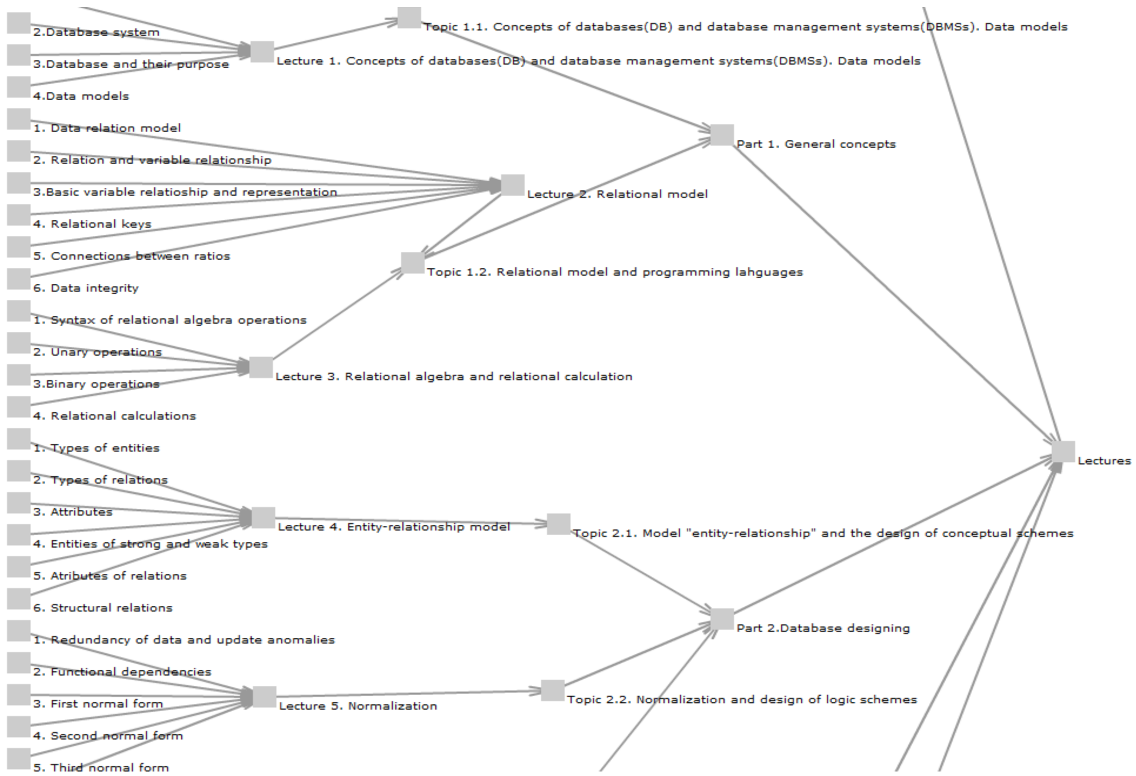


Figure 5b. Fragment of ontology to develop the credit module on discipline SOFTWARE OF DATA BASES (continued)

As one can see from the abovementioned example the whole educational standard and its course support could be represented in the form of aggregate of ontological models of the whole process of development of the bachelor's training course on the respective speciality. However, as one can see even from fragments of ontologies given in the respective Figures 1 - 5, we can make a conclusion that the axiom systems on each ontology and, moreover, functions of interpretation, do not match and may have not very robust overlapping. Thus, fragment of ontology SOFTWARE OF DATA BASES absorbs the whole axiomatics of ontograph SOFTWARE ENGINEERING. However the aggregates of interpretation functions of actions for each of the specified ontographs have both general and determined sets, which differ by semantics of elements.

To create the operational environment of ontological simulation of educational standards it is necessary to integrate the received models of component processes. Functional links between elements of aggregates, which determine notions-concepts, describe the certain procedures of process related to the patient's requests processing on problems of his/her health. Thus, we receive an aggregate, the elements of which are ontologies, which describe semantics of consultancy processes. The received aggregate of ontologies is determined as a single ontological model of interaction of processes related to resolving of consultancy tasks. Ontologies are combined and determined as a single ontological model of description of educational standard.

The separate formalized ontological models, which allowed to determining functions of interpretation of various levels of NFS, are based on an aggregate of functions of interpretation being specified on notions-concepts and their links.

Conclusion

Practical usage of ontological simulation requires application of various functions of interpretation. For instance, in our case the functions of determination of aggregate overlapping and analysis of matching are applied. Based on that it is necessary to apply the following procedure of integration of ontological models relates to processes of development and representation of various NFS levels that is designed on the basis of the following statement: aggregate of functions of interpretation of integrated ontology is not a combination of aggregates of functions of interpretations and ontologies of components, i.e., those which will combine.

Thus, ontological simulation of processes related to development of educational standards allows to optimizing a description of all HFS levels, determining a sufficient completeness of information content of subject courses, establishing correspondence between the systems of subject knowledge and competences, which should be developed among he graduates.

Additionally, the ontological models, due to their subject coherence, ensure correctness of development of inter-disciplinary links. It allows to optimizing the process of development of educational programmes, curricula and credit modules. Ontological models allow also to including into the curriculum the new technological solutions, knowledge of which is required for skillful specialist.

Bibliography

[Grossi, 2005] Grossi, Davide, Frank Dignum and John-Jules Charles Meyer. (2005). "Contextual Taxonomies" in Computational Logic in Multi-Agent Systems, pp. 33-51.

[Gruber, 1993] Gruber T.R. A translation approach to portable ontology specifications / T.R. Gruber // Knowledge Acquisition. – 1993. – Vol. 5. – P. 199 – 220.

[Guarino, 1994] Guarino N., The Ontological Level. In: Casati R., Smith N. and White G. (eds.), Philosophy and the Cognitive Sciences, Vienna: Holder-Pichler-Tempsky, 1994.

[Kleene, 1952] Stephen Cole Kleene,. Introduction to Metamathematics. North Holland. Aimed at mathematicians —.1952.– 560 p.

[Malishevski, 1998] Malishevski A.V. Qulitative models in the theory of complex systems. M.: Nauka. Fizmatlit. 1998, 528 c.

[Mendelson, 1997] Mendelson, Elliott (1997), Introduction to Mathematical Logic (4th ed.), London: Chapman & Hall. – 320 p.

[Rydeheard, 1988] D.E. Rydeheard, R.M. Burstall Computational Category Theory, New York: Prentice Hall, 1988, 257 p.

[Величко, 2009] Величко В.Ю. Автоматизированное создание тезауруса терминов предметной области для локальных поисковых систем / В. Величко, П. Волошин, С. Свитла // “Knowledge – Dialogue – Solution” International Book Series “Information Science & Computing”, Number 15. – Foi Ithea Sofia, Bulgaria. - 2009. – pp.24-31.

[Гладун, 1994] Гладун В. П. Процессы формирования новых знаний [Текст] / Гладун В. П. – София : СД «Педагог 6», 1994. – 192 с.

[Коршунова, 2009] Коршунова С. О. Роль тезаурусного моделирования в организации терминополя «ТЕХТ-ТЕКСТ»/ ВЕСТНИК ИРКУТСКОГО ГОСУДАРСТВЕННОГО ЛИНГВИСТИЧЕСКОГО УНИВЕРСИТЕТА - № 1, 2009. - http://cyberleninka.ru/article/n/rol-tezaurusnogo-modelirovaniya-v-organizatsii-terminopolya-text-tekst

[Палагин, 2005] Палагин А.В. Системная интеграция средств компьютерной техники / А.В. Палагин, Ю.С. Яковлев. – Винница: УНІВЕРСУМ, 2005. – 680 с.

Authors' Information

Oleksandr Stryzhak – *Institute of Telecommunications and Global Information Space of NAS of Ukraine, 186, Kyiv, Chokolivskiy blvd. 13; e-mail:* sae953@gmail.com

Major Fields of Scientific Research: Means for Knowledge Management, Transdisciplinary Ontology, Network Instruments, Transdisciplinary Integration of Knowledge Networks, Information Systems for Educational and Research Activities, Aggregation of Distributed Information Resources

KNOWLEDGE MANAGEMENT TOOLSET ANALYSIS FOR LEARNING ORGANIZATIONS

Kateryna Solovyova, Andrey Danilov, Panasovska Yuliia, Kobrin Maksim

Abstract: *The modern society's existence is inseparably linked with the development of information technologies. Today information plays a key role in the development of our society. As a result, science is beginning to move up to a more advanced stage of development - the noospheric one. The noospheric stage of scientific development involves the use of modern noospheric system approach and innovative methods and technologies based on knowledge. This work is based upon a noospheric system methodology (systemology) which allows to consider complex systems holistically, to take into account their essential (ontological) properties and relations, to solve the complicated ill-structured problems relying on knowledge. The application of the systemological classification analysis on the basis of the natural classification criteria allows to define and to simulate deep knowledge being adequate for any subject domain including the ill-structured ones; to consider the objects' essential properties. Using of the knowledge-oriented system tools allows employees to build intellectual capital, will allow organizations and states to develop in a more efficient, competitive way and to form knowledge economy. Knowledge management tools are the most effective means of managing organizations to improve their competitiveness.*

Alongside the increasing number of Internet users, the enhancing of its availability, the effectiveness of social networking on the Internet in organizations rises, it is especially vital for employees training. The improvement of quality of social networks functioning on the Internet requires the adequacy of their domain, adaptability and user-oriented interface. The paper has presented some results of studies on knowledge systematization in the subject domains "Social Networking on the Internet", "Competences" on the basis of systemological classification analysis that is directly relevant for the education field, the learning organizations creation, functional platform for further development of this direction.

Keywords*: knowledge management, a learning organization, a social network on the Internet, a competence, knowledge-oriented technology, the Internet, systemological classification analysis, classification, a systemology model, education, noospheric stage of scientific development.*

Introduction

The enhancing of information technologies' role in society is a natural process that has accelerated within the recent years. The emergence of new methods, techniques and approaches to the various subject domains' activities requires the use of modern information tools that will significantly improve the employees' efficiency.

The modern information technologies application is of significant importance to the rapidly growing industries as well as to the industries which depend on the society's functioning and development. The life protection, environment industries etc. are of great consequence. But there is a direction which is not possible without a sustainable development of the company, organization, person, government and society as a whole; and that is knowledge-based direction, knowledge management, and learning.

To improve the education quality and training level it is necessary to meet the scientific development trends and educational methods. The use of modern teaching methods, knowledge management elements particularly will significantly improve its potency and effectiveness. The current developments in the knowledge management field corresponding to the noospheric stage of scientific development have been considered in the paper.

The object of the research

The object of the research is to present the knowledge-based systemological approach within the "Social Networking" and "Competences" domains, aimed at improving the education quality and the organization's intellectual capital level.

The method of the research

As a method of the research and solving the set tasks, the systemological classification analysis method, which refers to the noospheric stage of scientific development and is able to effectively solve ill-formalized problems, has been employed. The diagram presenting the systemological classification analysis principal steps has been given iteratively at **Figure 1**.

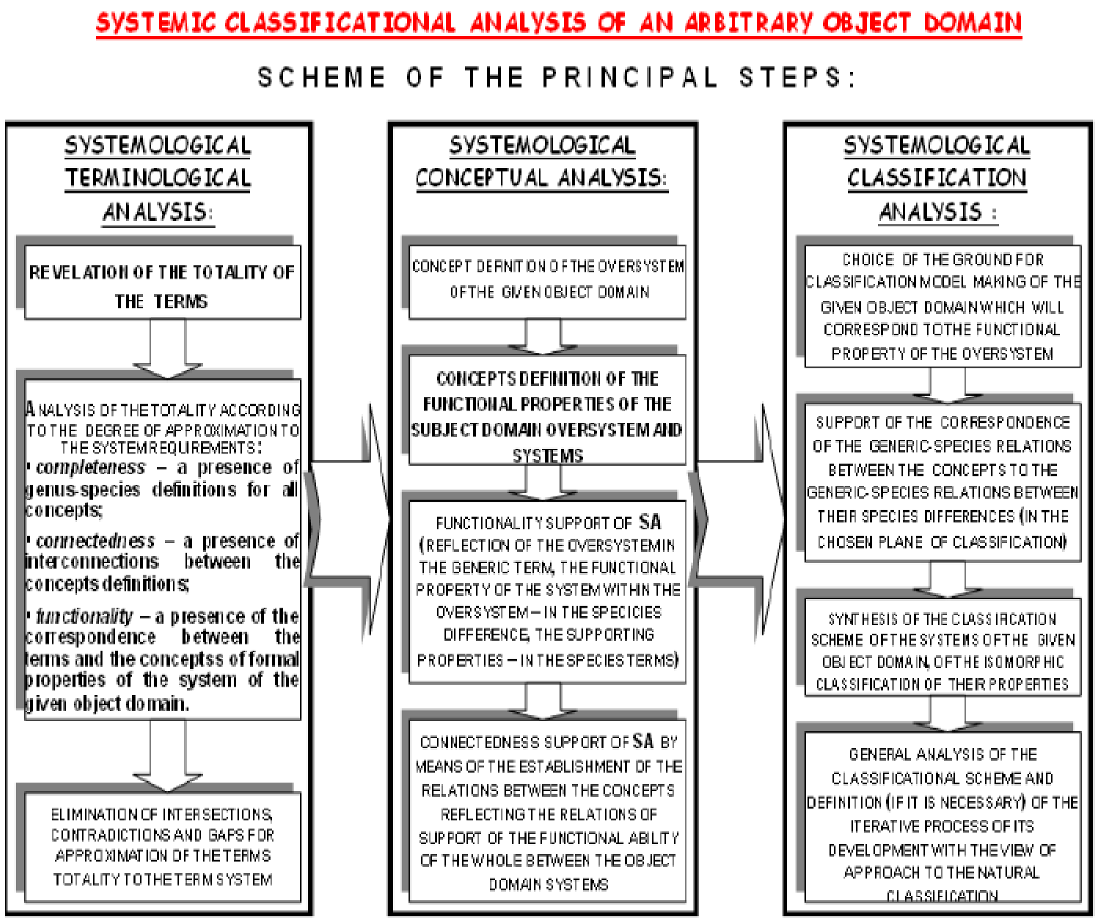


Figure 1. The scheme of the principal steps of systemological classification analyses [1]

The systemological classification analysis method application allows us to construct and evaluate any classification (taxonomy) in terms of its validity (parametricity), domain's adequate reflection and objects' essential properties. The method also enables the objects' detection and prediction depending on their location and properties in the classification and allows using classification as a theoretical analysis tool in the relevant domain and successful implementation.

The social networks development urgency in the educational sphere

The information technologies development leads to the necessary application of the up-to-date business practices and information professionals' training. An increasing number of different human activity sectors integrate their functions in the information sphere; nowadays it is very difficult to imagine any human activity area that does not use information technologies and does not depend on the Internet in their daily work. This is especially true of the state strategic sectors which determine the society development to the great extent. One of such industries is education through which companies and organizations can get modern highly qualified specialists [2].

Social networks provide the educational system with many potential means:

- Easy exchange of experience and knowledge between people from different cities (countries) without the personal presence that will bring both foreign students and professors from leading universities in the world to educational and scientific processes;
- Access to the network members' information. Using a university's social network will provide large amounts of information on scientific subjects which will be classified basing on the systemological classification analysis method; this will greatly reduce the seek time and increase the useful scientific work amount;
- Fast and easy way to get advice from experts and subject area;
- Feeling of being involved into universities' innovations, ideas and current developments which are produced as a result of discussions.

To develop effective social networks that will successfully fulfill the set tasks it is necessary to apply modern scientific methods oriented to the work with the essential aspect of information. The most effective modern method of complicated ill-structured problems solution is the systemological classification analysis method relevant to the noospheric stage of scientific development [2].

Using of the systemological classification analysis in functions classifications construction allows to work out recommendations for the functions arrangement in a social network's menu as for their meaningful arrangement according to the classification developed. This natural arrangement will significantly reduce the user' load and improve his/her work [2].

The proposed social networking application in the field of distance learning involves not only training materials creation in the Internet environment and the individual work of students, but also the opportunity of direct contact between teachers and students at any time. This approach requires a significant increase teacher workday due consultation at a convenient time for students. This approach to the disciplines study allows organizing the training in the form of a creative workshop, and that will greatly enhance the learning effectiveness through direct contact between teachers and students [2].

Knowledge-oriented social networks application in the learning process will help to organize tutoring students without their personal presence at the university, which will contribute to better employment, improve the training quality and will allow people with special needs to receive qualitative education in a comfortable environment [2].

Analysis of learning organizations as a form of the organization of business

The emergence of a learning organization is considered as a natural reaction to the conditions of the transition from the information economy to the knowledge economy, in which there is a need for new organizational structures and tools. The unique feature of such an organization is not only the staff possessing their common experience and knowledge, but the culture of the organizational learning and development. The staff apprehends their job as a process open to continuous improvement [3].

"A learning organization is an organization that generates, obtains and distributes knowledge constantly and continuously, changes and improves its behavior depending on its own experience learning, creates new products and services, permanently using the employees' ideas and analyzing the customers 'and partners' knowledge. As it can be seen from the definition, a learning organization is based on continuous control of its intellectual capital" [4].

The most famous model of a learning organization belongs to Peter Senge, this model is based upon five "skills" or "disciplines", which each employee of the organization should perfect himself/herself in, so that it becomes a true learning [5]: personal skills, mental models, common vision creating, group learning, system thinking.

For a clear and unambiguous understanding of the knowledge management system implementation process the standard based on the practical guide of knowledge management in Europe "Knowledge Management. Terms and definitions" [6] has been worked out, it is aimed at all business members, implementing a knowledge management system, both within individual companies and in their cooperation. Knowledge management is one of the least studied modern management structures. Increasingly, the problem of the knowledge management theory becomes a subject of economic debate, due to the increasing role and influence of knowledge as an economic resource for economic activities of different levels.

One of the main challenges of knowledge management is the clear definition of the terms and concepts and their systematization. Organizations (especially small and medium-sized enterprises) may undergo a significant economic impact, using common terminology and definitions. Therefore it is especially important to use systemological classification analysis, which allows to systematize conceptual knowledge in the most objective and justified way.

It is also advisable to use a good taxonomy (parametric classifications close to the natural ones) in learning process, and systemological classification analysis itself as a methodological framework and system tool to work with knowledge. It also will contribute to the development of organizations' stuff and their competences.

The competences analysis for learning organizations

Staff development is one of the most important activities of stuff management and successful production factors. At the same time investing in staff development plays a greater role than investing in the production facilities development and improvement. Under staff development we understood a set of measures aimed at advanced training and employees' psychological characteristics improvement. This refers primarily to [7]:

- Training that provides the necessary knowledge, skills and experience in the form of general and vocational education;
- Training the mission which is to improve professional knowledge and skills;
- Advanced training which is defined as any professional reorientation. The purpose of advanced training is to enable employees to learn a new profession for them.

Education (training, seminars, distance learning, coaching, etc.) is always directed at the competences required to perform the tasks and achieve the desired results [8].

Formation of skills is the result of skillful use of different teaching methods altogether with the specific situation, values, skills and knowledge. For example, the successful management of the team may be the result of effective teaching methods such as interviews to evaluate the performance, workshops, providing "feedback", performance management [9]. Formation of skills is a complex process consisting of several stages.

At the first stage of formation of competences it is necessary to determine the competences themselves, the formation of which the organization requires on the basis of the system determination of the organization's mission.

When the competences are defined, it's time to train relevant knowledge and skills at the employees as well as to develop existing skills. It is necessary to determine the training methods and programs, as well as employees who are interested in the formation of a particular skill. Then you can proceed directly to the training and as a result, the formation of competences.

A competences formation model using the IDEF0 standard and the current system approach has been developed. Under the three-level representation of the model the system approach has been taken into account.

The model consists of four principal stages of the competences formation:

- Creating of a competences model involves creating a list of skills necessary to be formed at the organization in general;
- Creating of occupational skills profiles includes the process of identifying competences necessary to form at each employee individually;
- Corporate training includes training staff to develop the necessary competences;
- Checking learning outcomes includes examination to check the knowledge received by employees in the course of training.

The application of the systemological classification analysis and system approach in the competences formation can facilitate and accelerate this process. This in turn leads to rapid development and competitiveness of the organization.

An example of the systemological classification analysis application

The analysis of the sources revealed that "competence" is a complex, multicomponent, interdisciplinary concept defined by many different definitions. The descriptions of this concept's idea are lax and differ in size, structure, semantic and logical structure, and not all of them are of generic character.

Specialization "Competences" is ill-structured and has no standard classification. However, the classifications according to different parameters such as the sharing competences level, the development the level, the essence and content, the emergency method etc. are possible.

To create a classification of the "Competences" subject domain and the detailed consideration of "Professional Competences" the systemological classification analysis method has been employed, so we have chosen "Function Purpose" as a basis for our classification (Figure 2).

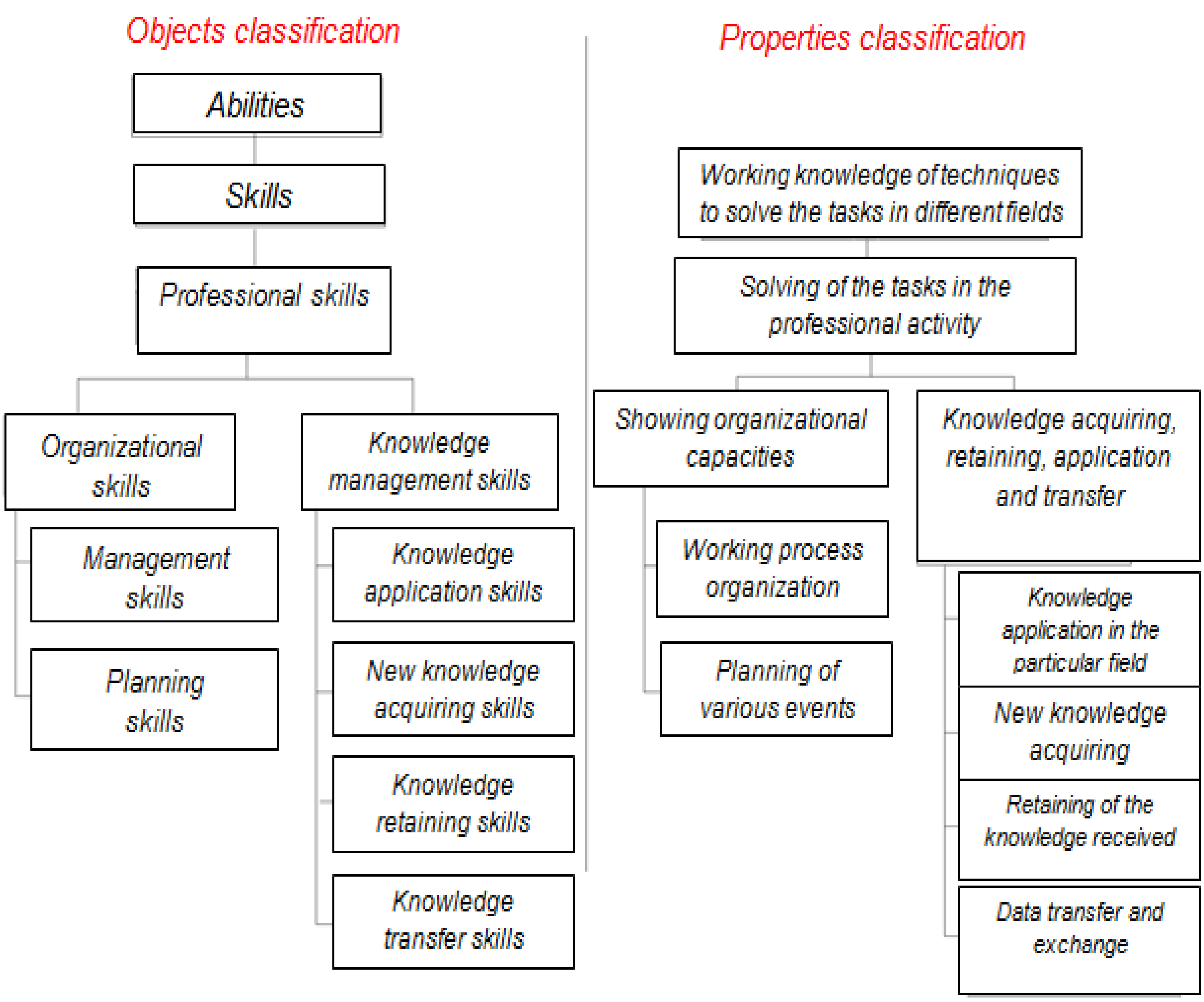


Figure 2, A Fragment of the "Competences" Classification

The systemological cognitive method and knowledge management application is useful in various fields and improves problem solving, for example in the field of information security management, without which the successful development of any organization is not practically possible. In some companies, as yet, the knowledge management system is a sphere of activity of the HR management and IT department employees (activity provides the technical possibility of knowledge management system). The potential of the knowledge management system is not used to the full, knowledge, human factor (which affects the safety significantly)and employees' cognitive features are not taken into consideration in a proper way, information security professionals are not integrated enough into the work of the knowledge management system. Therefore, the issue of knowledge management related to the work of information security systems and the knowledge-oriented systemological cognitive approach application are pressing.

The systemological classification analysis method application and employees' cognitive features consideration as a bottleneck in the information security system are the effective way to solve the given task and result in the competitive advantage achievement.

The findings

In the result of the conducted work the following findings has been obtained:

- The peculiarities, relevance and effectiveness of the knowledge-oriented system methods and techniques in the field of learning organizations and intellectual capital have been analyzed;
- The analysis of learning organizations has been held;
- The main aspects of niche social networks construction on the Internet based on system logical cognitive approach to improve the quality of the educational process have been studied;
- The classification of competences on the basis of the system logical classification analysis has been studied and developed; it is also used in the employees' competences formation modelling to increase the organizations' intellectual capital and to accelerate the learning process;
- The system logical classification analysis application allows systematizing knowledge reasonable and adequate for their subject domain, including ill-structured ones; to receive deep knowledge, to create parametric knowledge classification with objects' essential (ontological) properties.

The research in the considered subject area will allow implementing the developed models into the learning process effectively in order to improve the quality of training at learning organizations inter alia.

Acknowledgements

The paper is partially financed by the project **ITHEA XXI** of the Institute of Information Theories and Applications FOI ITHEA and the Consortium FOI Bulgaria (www.ithea.org, www.foibg.com).

References

1. Соловьева Е.А. Естественная классификация: системологические основания / Е.А. Соловьева. [Текст] – Харьков: ХНУРЭ, 1999. – 222с.

2. Данилов А.Д. К вопросу об актуальности разработки знаниеориентированных социальных сетей в сфере образования. / 18-й Международный молодежный форум «Радиоэлектроника и молодежь в XXI веке». Сб. материалов форума. Т.9. Международная конференция «Управление знаниями и конкурентная разведка» - Харьков: ХНУРЭ. 2014. – с.50 -51.

3. Обучающаяся организация [Электронный ресурс]. – Режим доступа: http://www.science-education.ru/106-7407 – 10.05.2013 г. – Загл. с экрана.

4. Концепция интеллектуальной организации [Электронный ресурс]. – Режим доступа: http://www.m-economy.ru/art.php?nArtId=1507 – 10.05.2013 г. – Загл. с экрана.

5. Концепция интеллектуальной организации [Электронный ресурс]. – Режим доступа: http://www.znannya.org/?view=learning-organization – 10.05.2013 г. – Загл. с экрана.

7. Развитие и профессиональное обучение персонала [Электронный ресурс] - Режим доступа: www/URL: http://www.bibliotekar.ru/biznes-33/23.htm – 15.05.14 – Загл. с экрана.

8. Использование компетенций для тренинга и развития [Электронный ресурс] - Режим доступа: www/URL: http://www.smart-edu.com/ispolzovanie-kompetentsiy-dlya-treninga-i-razvitiya. html – 15.05.14 – Загл. с экрана.

9. Компетенции персонала в современной организации [Электронный ресурс] – Режим доступа: www/URL: http://www.e-xecutive.ru/knowledge/announcement/694102/ – 10.04.14 – Загл. с экрана

Authors' Information

Kateryna Solovyova *- Chief of Social Informatics Department and Knowledge Management Center, Professor, Doctor of Technical Sciences, Kharkov National University of Radio Electronics, Lenin Ave., 14, Kharkov, 61166, Ukraine; e-mail: si@kture.kharkov.ua.*

Major Fields of Scientific Research: Knowledge Classification, Systematization, Elicitation, Acquisition and Modeling, Knowledge Management, Ontological Engineering, Systemological Analysis, Knowledge Research and Application, Decision Making, Knowledge-Based Systems and Technologies, Artificial Intelligence, Business Intelligence, Modern (e-) Learning, Competitive Intelligence, Cognitive Modeling, Intellectual Capital .

Andrey Danilov *– Social Informatics Department, Kharkov National University of Radio Electronics, Lenin Ave., 14, Kharkov, 61166, Ukraine; e-mail: si@kture.kharkov.ua, Skil06@ukr.net.*

Major Fields of Scientific Research: Social Networks, Ontological Engineering, Competitive Intelligence, Decision Making, Intelligence Technologies, Knowledge Research and Application, Knowledge Management, (e-) Learning, Artificial Intelligence, Systemological Analysis, Social Capital.

***Panasovska Yuliia** – MA Student of Social Informatics Department, Kharkov National University of Radio Electronics, Engineer, Lenin Ave., 14, Kharkov, 61166, Ukraine; e-mail: si@kture.kharkov.ua. Major Fields of Scientific Research: Ontological Engineering, Intelligence Technologies, Knowledge Research and Application, Knowledge Management, Systemological Analysis.*

***Kobrin Maksim** – Sotial Informatics Department, Kharkov National University of Radio Electronics, Lenin Ave., 14, Kharkov, 61166, Ukraine; e-mail: si@kture.kharkov.ua, infoma@rambler.ru.*

Major Fields Scientific Research: Knowledge Management, Artificial Intelligence, Information Security, Systemological Analysis, Intelligence Technologies.

AN AGENT-BASED CONCEPT FOR PROBLEM MANAGEMENT SYSTEMS TO ENHANCE RELIABILITY

Huiqiang Wang, Nasser Jazdi, Peter Goehner

Abstract: *A defective component in an industrial automation system affects only a limited number of sub functions. However, the affected sub functions often lead to break down of the whole system. The goal of this paper is to propose an intelligent agents-based concept for a Problem Management System (IAPMS) to enhance the reliability of industrial automation systems by resolving not only the known problems but also the unknown problems. In this paper, six types of component agents represent all the necessary components of an industrial automation system and a coordination agent type orchestrates the coordination between the component agents. The component agents detect the defective component by using the existing diagnosis system. On one hand, the component agents and the coordination agent could solve the known problems by using the pre-set measures. On the other hand, since no measures for an unknown problem exist, the corresponding component agents negotiate with the coordination agent to keep the unaffected sub functions alive; hence the system can be further used.*

Keywords: *reliability, intelligent component agent, Problem Management System, unknown problems*

ACM Classification Keywords: *I.2 Artificial Intelligence – I.2.1 Applications and Expert Systems – Industrial automation*

Introduction

Nowadays, usability, reliability and stability of the industrial automation systems are getting more important. To prevent and to deal with the possible problems, the developer will perform many measures in the industrial automation systems. However, some unknown problems may occur during the operation phase as reported in [Baskerville, 2004; Luguna, 2003; Kim, 2002; Mockus, 2000; [Musa, 1993]. The reasons are concluded as followed:

- Short development time and constraint of budget;
- No systematic reuse of the software;
- Resource constraints in test;
- Defective experience of developer;
- Not completely known operational profile in the development phase.

Meanwhile, there are no existing solutions for the emerging problems, this type of the problems are referred to as ***Unknown Problems***. In this case, a component of an industrial automation system can either not properly be performed or cannot be executed at all, as well as its consequent sub functions. This can lead to malfunctions of the whole system. In most cases, a sub function of the system can also meet the users' requirements. This is to say, because of the unknown problems, the reliability of the unaffected sub functions is also declined.

In this paper, we propose an agents-based concept for Problem Management System that attempts to deal with the unknown problems and to perform the unaffected specific sub functions, in order to enhance the reliability of the specific functions. All of the components in an industrial automation system will be represented by intelligent component agents, i.e. user agent, computational unit agent, sensor agent, bus agent and actuator agent. Bus agent type consists of field bus agent and communication agent. It is necessary to have a coordination agent for the Problem Management System, in order to find proper solutions for the unknown problems. Furthermore, a component agent realizes the problem in its component and attempts to solve it by itself. Afterwards, the component agent consults with the coordination agent on the unknown problems. In the end, they succeed in finding solutions to solve the problems, as well as preforming the unaffected sub functions, thereby improving the reliability of the sub functions.

This paper is organized as following: section 2 introduces the theoretical aspects of an industrial automation system and reliability, and state of the art refers to the approaches of problem management in an industrial automation system. A comparison among the approaches will be performed. In section 3 we will show the architecture of agents-based Problem Management System. In section 4, we present the workflow of problem management in the concept. In section 5 we give some possible applicable scenarios of IAPMS. Finally, conclusions are drawn and provided (Section 6).

Theoretical Aspects and State of the Art

Reliability in an industrial automation system

An industrial automation system consists of user, computational unit (CU), actuators and sensors [Sharma, 2011]. The following part describes the different components of an industrial automation system:

- User: inputs value to the computational unit by employing the user interface;
- Computational Unit: achieves the tasks that are assigned by the user. The input value will be processed and set to the actuators;
- Actuator: receives commands from the computational unit and influences the technical plant;
- Sensor: measures the physical values of technical plant and converts it into electrical values for the computational unit;
- Bus: Depending on the size of the system might consist of multi levels. Typically, in a plant automation we have field bus and plant bus. Field bus transports the signals between the computational unit and the actuator, as well as the sensor; Plant bus transports the signals between the user and the computational unit.

Reliability according to IEEE 610.12-1990 [IEEE, 1999] is defined as the ability of a system or a component to perform its required functions under stated conditions for a specific period of time. This definition consists of four parts: ability, required functions, conditions, time. The required functions are such functions that are specified by the customer or the developer.

In an industrial automation system, the different components impact the different sub functions and the overall function consists of a number of sub functions. When a component in an industrial automation system is defective and there are no solutions for the occurred problem, the corresponding sub functions will be break down, leading to the stop of the whole system. However, the unaffected sub functions, which can fulfill the requirement of the user, could not be performed because of the loss of the overall function. So we propose a definition of ***Reliability of the sub functions***: the ability of an industrial automation system to perform its required sub functions under stated conditions for a specific period of time.

In order to perform the required functions, some measures must be preset to figure out the possible problems in the operation phase. In our paper, the ***Unknown Problem***, which can be solved by the agents-based concept of Problem Management System, is defined as followed:

- A defective component could lead to the stop of a whole industrial automation system;
- The cause of the defective component is unknown;
- There are no solutions to solve the occurred problem;
- The defective component can be detected by an existing industrial automation system or an existing fault management system [Bordasch, 2013].

Problem Management in an industrial automation system

To manage the occurred problems or faults in an industrial automation system, four types of conventional approaches are regrouped in the following four techniques [Lussier, 2004].

Fault prevention: this approach attempts to prevent the occurrence or introduction of faults. Bordasch [Bordasch, 2013] proposes to prevent a fault using a functional model and a hybrid abnormity identification concept. In this approach, all abnormities are identified and diagnosed. Then the system generates the removal actions and assists the technical staff to solve the problem. This approach aims to sustain the system in a fault-free state. However, in case of an unknown problem, i.e. a problem that has never had occurred, no solution can be provided.

Fault removal: this approach tries to reduce the number or severity of faults. The work by [Stutzke, 2001] proposes a stochastic model that relates the software failure intensity function to development and debugging error occurrence throughout all software life-cycle phases, in order to remove the failures in the development phase. This approach aims to develop a system without faults. Yet since no measures for detecting unknown problems exist, this approach cannot solve the unknown problems in an industrial automation system.

Fault tolerance: this approach attempts to deliver correct service in the presence of faults. Generally, it composes of error detection and subsequent system recovery. For the problem, the error handling is performed. It can take three forms: rollback, roll forward, redundancy. Redundancy is usually used for fault tolerance in an industrial automation system [Kauer, 2014]. Fault tolerance aims at the occurred errors or problems, but in the practical field, it's almost impossible to fulfill the redundancies of all the components in an industrial automation system.

Fault forecasting: this approach attempts to estimate the present number, the future incidence and the likely consequences of faults. This approach can forecast the possible faults. But when there are no corresponding measures, the problem will not be solved.

As it can be concluded, it is impossible to solve the unknown problems by the conventional approaches in an industrial automation system. Modern research activities in the field of intelligent agent systems has opened new horizon for managing the faults in an industrial automation system. In [Merdan, 2011], Merdan presents an automation agent approach with agents comprising a software component with an integrated world model repository besides the related hardware. World model is used for the representation of the external surroundings and internals of the agent. In [Cerrada, 2007], Cerrada proposes a reference model for fault management in industrial processes. A set of models describe the general characteristics of the agents, specific tasks, communications and coordination. In the fault management system, the actions are related to the decision-making in the scheduling of the preventive maintenance task and the running of preventive and corrective specific maintenance tasks. But there are no approaches to perform the unaffected functions. Consequently, a new concept for the Problem Management System is proposed in this paper. With the help of the intelligent agents, the Problem Management System provides a possibility to analyze the occurred problems, to search the unaffected sub functions, and to make a new decision by itself. Furthermore the Problem Management System can help an industrial automation system to perform the unaffected sub functions as well as to maintain the reliability of the sub functions.

Architecture of Agents-based Problem Management System

In the last section, we have reviewed some conventional approaches without agents and some approaches with agents to solve the problems in an industrial automation system. In this section, we propose a new architecture of agents-based Problem Management System to enhance the reliability of the sub functions (IAPMS). Hence, we use different component agents to represent different types of the system components by using the abstraction of knowledge, information and data of the components in an industrial automation system.

Component agents

In IAPMS, the component agents represent all components in an industrial automation system. They consist of seven types of agents (see Figure 1):

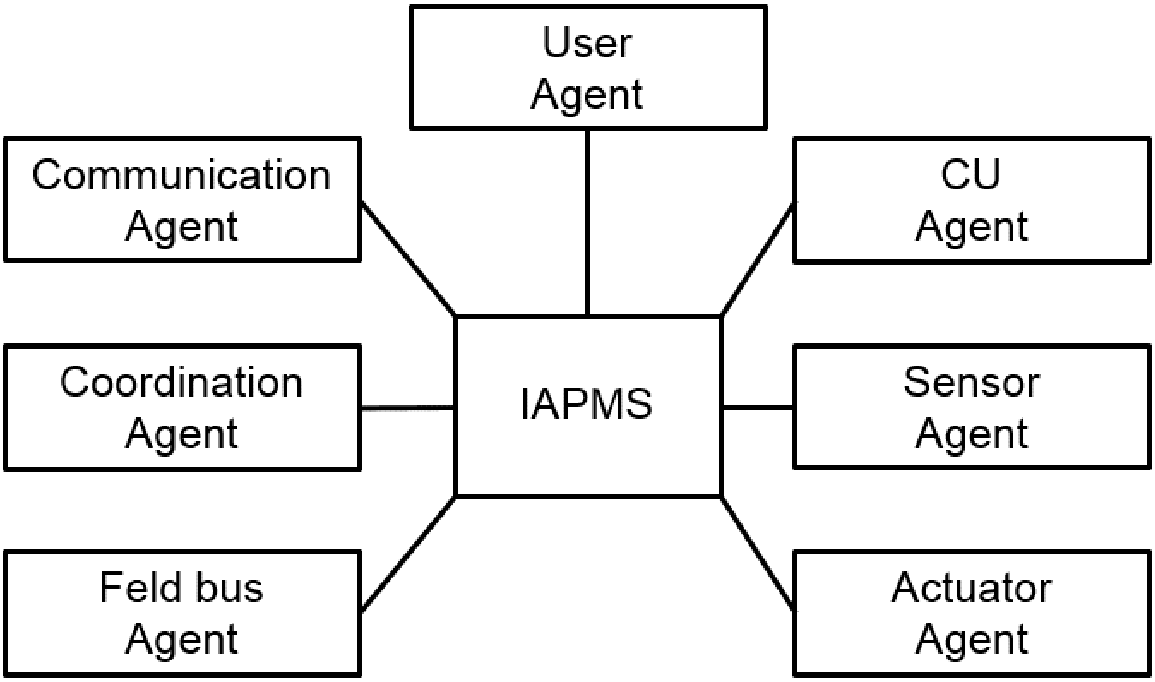


Figure 1. Six types of component agents and a coordination agent in IAPMS

- Sensor agent: It represents all the sensors in an industrial automation system. It knows the defective sensors and the affected sub functions. Furthermore, the sensor agent can perform the pre-set solutions or negotiate with the coordination agent.
- Actuator agent: It represents all the actuators in an industrial automation system. It knows all the relationships among the actuators and the affected sub functions. With a defective actuator, the actuator agent can perform the pre-set solutions or negotiate with the coordination agent;
- Computational unit agent (CU agent): It represents all the computational units in an industrial automation system. It knows the processes or domains of the computational unit and the corresponding sub functions as well as the sub-systems;
- User agent: It represents the user of an industrial automation system. It knows the task and the required parameters of the user and it is in the charge of communication between the user and the coordination agent. Moreover, the user agent could provide an interface for the user [Mubarak, 2010];
- Field bus agent: It represents the field bus between the computational unit and actuators as well as sensors;
- Communication agent: It represents the communication between the user and the computational unit;
- Coordination agent: a coordination agent is needed to find a solution for the problems by coordinating the component agents. Coordination agent coordinates the agents' community [Cerrada, 2007] and deals with the unknown problems in an industrial automation system. It negotiates with the component agents to make a new decision to deal with the defective component and to perform the unaffected sub functions.

The Architecture of IAPMS

IAPMS includes six types of component agents and a coordination agent (see figure 2). The intelligent component agents represent different types of system components in an industrial automation system. The occurred problem of a component will be firstly handled by the corresponding component agent. The coordination agent can communicate with the corresponding component agent and make a decision to handle the unknown problems. The user agent is a special component agent. It represents the interface between a user and an industrial automation system and deals with the possible problems in the interface. Moreover, the user agent provides an interface for the user in the IAPMS. The user can decide whether to perform the decision by the coordination agent or not. The database stores a lot of decision for the different problems. With the help of the cloud, the coordination agent can remote access through the internet the resources in the server. It is necessary to have the required information on the server. With this the IAPMS can get further information, e.g. similar problems within other industrial automation systems, and employ them for better decisions.

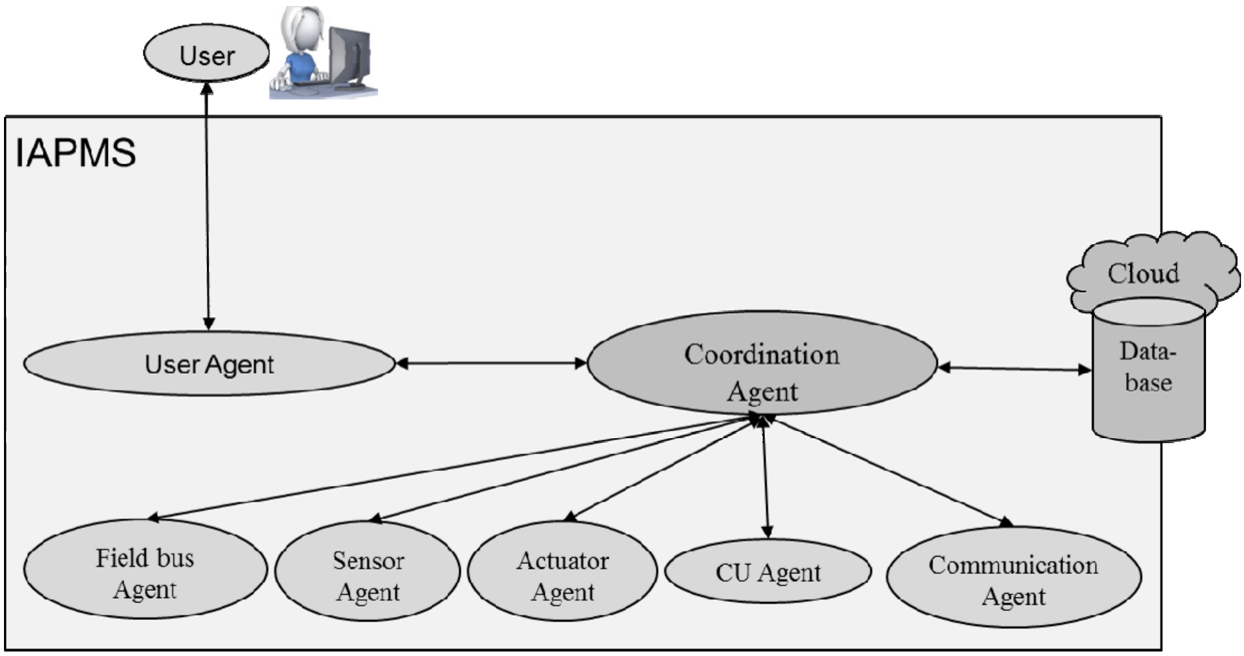


Figure 2. Architecture of IAPMS

Workflow of IAPMS

In this section, the workflow of IAPMS will be illustrated in figure 3. The workflow of the problem management consists of four parts: realizing the problem, dealing with the local problems, dealing with global problems, and communicating with database.

Realizing the problem

Firstly, when a problem occurs, the component agent can realize via the existing diagnosis system that there is a problem in its represented component, i.e. the corresponding sub function cannot be fully performed. The component agent identifies if the problem has occurred only in its component, i.e. whether the problem is local or global. A local problem means that the problem affects only one component; a global problem means the problem affects more than one component.

Dealing with local problems

If the occurred problem is local, the component agent searches its knowledge whether there is a known solution to the problem. If there is a solution, the IAPMS implements the solution (either solves the problem or provides a restricted functionality). If the solution is insufficient, IAPMS informs the maintenance service; If there is no known solution, the component agent negotiates with the coordination agent to find a solution to handle the problem (e.g. perform the not affected sub functions, which can fulfill the requirement of the user). If after the negotiation, the coordination agent cannot find a solution, then the IAPMS will again inform the maintenance service.

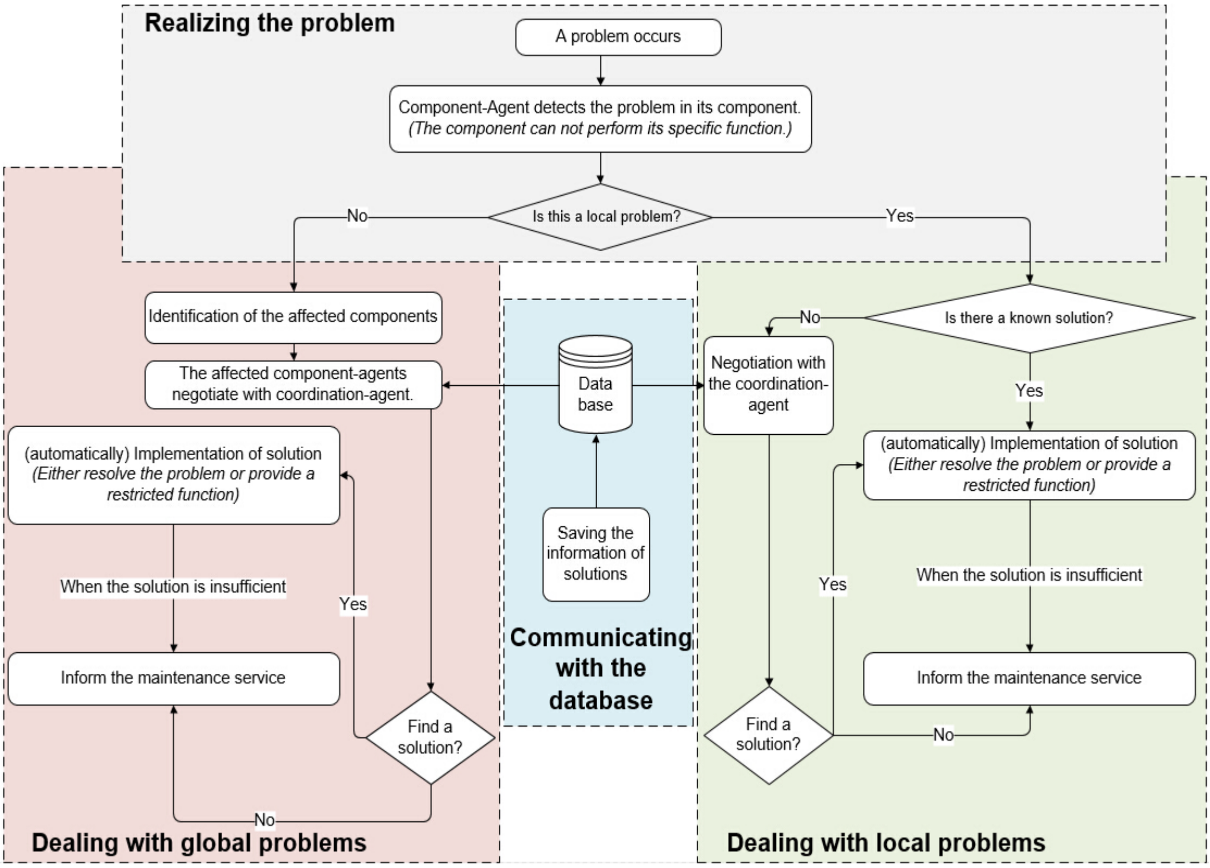


Figure 3. Workflow of IAPMS

Dealing with global problems

If the occurred problem is global. The component agent should identify all the affected components. All the affected component agents negotiate with the coordination agent to find a solution. In addition, the coordination agent can remote access the resource on the server via internet. If a solution is found, then the IAPMS should implement the solution (e.g. perform the unaffected sub functions, which can fulfill the requirements of the user or the customer). If after the negotiation, the coordination agent cannot find a solution or the solution is insufficient, then the IAPMS will inform the maintenance service.

Communicating with the database

When the coordination agent is needed to solve the problem, the coordination agent inquiries the database, which stores the solutions for the known problems and the entire sub functions of an industrial automation system. After the negotiation among the coordination agent and the component agents, the solution should be stored in the database by the coordination agent. Hence, the solution can be used for the next time.

Possible Applicable Scenarios of IAPMS

For demonstrating the possible applicable scenarios, the applications of intelligent agent-based Problem management System will be introduced in this section.

Scenario 1: Industrial Coffee Machine

Problem: In an industrial coffee machine, after several dispensing of coffee, the drip tray of the coffee machine is full (see figure 4). Then the coffee machine will stop working unless someone empties the drip tray. Because of the full drip tray, the sub function "hot water" is also out of order. However, in the real operation phase, someone may just need the sub function "hot water". While this problem is not considered in the development phase, the coffee machine has no solution to solve this unknown problem in the operation phase and to fulfill the requirement of the users.

Solution: On this occasion, all the components in the industrial coffee machine are represented by component agents, and the sensor agent can detect and analyze the problem. By the inquiry of the database and the reasoning of all the corresponding component agents, a new solution can be reasoned by the coordination agent to perform the sub function hot water.

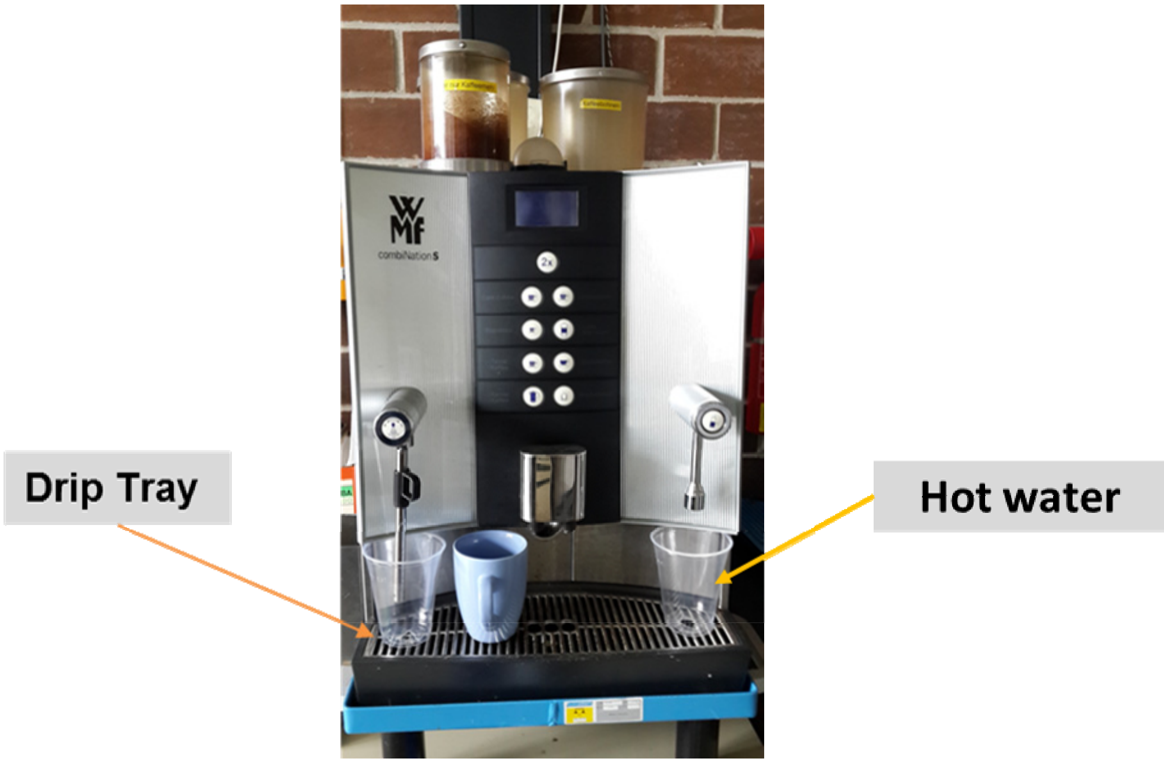


Figure 4. IAS Industrial Coffee Machine

Scenario 2: High-bay Warehouse

Problem: Here we consider a model of a high-bay warehouse at the Institute of Industrial Automation and Software Engineering. It consists of three sub functions, input the work pieces from the transporter, store the work pieces in a slot, and output the work pieces to the users. Figure 5 shows a picture of the high-bay warehouse at IAS. Because of the requirement of the users, typically, the direction from input to output is fixed. However, when the motor of input is broken, the high-bay warehouse must stop to wait for the maintenance service, until the motor has been repaired. In this case, the sub function of storing and output is broken down.

Solution: On this occasion, all the components in the high-bay warehouse are represented by the component agents. The broken motor can be tested by the existing industrial automation system. Then the actuator agent realizes the problem. Because there are no existing solutions in the actuator agent and its knowledge, the actuator agent must communicate and negotiate with the coordination agent to find a solution as well as to perform the sub function of storing the work pieces and outputting the work pieces.

Figure 5. IAS High-bay Warehouse

Conclusion and Future Work

In this work, we reported an intelligent agents-based approach for Problem Management System (IAPMS) to enhance the reliability of specific functions. This approach improves not only the reliability of the whole system but also improves the reliability of the sub functions of an industrial automation system. Meanwhile, a novel way is proposed to deal with the occurred unknown problems by performing the unaffected sub functions that might still fulfill the requirement of the user.

For the realization of the concept, all the components in an industrial automation system will be represented by the component agents. In the operation phase, a defective component affects only a limited number of sub functions, the affected sub functions result in the stop of the whole industrial automation system. The component agent realizes the defective component through the existing diagnosis system. The component agents and the coordination agent solve the local problems by pre-set measures. Afterwards, since no solutions for a problem, the corresponding component agents negotiate with the coordination agent to keep the unaffected sub functions alive, which can still fulfill the requirement of the user. In addition, the coordination agent can access the resources on the server via Internet to find a possible solution, assuming that the necessary information are available.

As a future work, we are implementing the concept of IAPMS in a practical project, on an industrial coffee machine and a high-bay warehouse at IAS. Evaluation of the results based on the prototype will follow.

Acknowledgements

We thank Chinese CSC (China Scholarship Council fellowship Grant) for financial support.

Bibliography

[Baskerville, 2004] R. Baskerville and J. Pries-Heje. Short cycle time systems development. In: Information Systems Journal, vol. 14, no. 3, pp. 237-264, 2004.

[Bordasch, 2013] M. Bordasch and P. Goehner. Fault prevention in industrial automation systems by approaches of a functional model and a hybrid abnormity identification concept. In: Proceedings of the IECON 2013-39th Annual Conference of the IEEE, pp. 2845-2840, IEEE, 2013.

[Cerrada, 2007] M. Cerrada and et al. Agents-based design for fault management systems in industrial processes. In: Computers in Industry, vol. 58, no. 4, pp. 313-328, 2007.

[IEEE, 1999] Institute of Electrical and Electronics Engineers, IEEE Std. 610.12-1990, IEEE Standard Glossary of Software engineering Terminology, 1999.

[Kauer, 2014] M. Kauer and et al. Fault-tolerant control synthesis and verification of distributed embedded systems. In: Proceedings of the Conference on Design, Automation and Test in Europe, pp. 1-6, European Design and Automation Association, 2014.

[Kim, 2002] J. Kim and A. Porter. A history-based test prioritization technique for regression testing in resource constrained environments. In: Software Engineering, 2002. ICSE 2002. Proceedings of the 24rd International Conference on, pp. 119-129, 2002.

[Luguna, 2003] M. Luguna and et al. Introducing systematic reuse in mainstream software process. In: in Proceedings of the Euromicro Conference, pp. 351-358, IEEE, 2003.

[Lussier, 2004] B. Lussier and et al. On fault tolerance and robustness in autonomous systems. In: Proceedings of the 3rd IARP-IEEE/RAS-EURON Joint Workshop on Technical Challenges for Dependable Robots in Human Environments, pp. 351 -358, laas.fr, 2004.

[Merdan, 2011] M. Merdan and et al. Monitoring and diagnostics of industrial systems using automation agents. In: International Journal of Production Research, vol. 49, no. 5, pp. 1497-1509, 2011.

[Mockus, 2000] A. Mockus and D. Weiss. Predicting risk of software changes. In: Bell Labs Technical Journal, vol. 5, no. 2, pp. 169-180, 2000.

[Mubarak, 2010] H. Mubarak and P. Goehner. An agent-oriented approach for self-management of industrial automation systems. In: Proceedings of the Industrial Informatics (INDIN), 2010 8th IEEE International Conference on, pp. 721-726, IEEE, 2010.

[Musa, 1993] J. Musa, Operational profiles in software-reliability engineering, vol. 10. IEEE, 1993.

[Shannon, 1949] C.E.Shannon. The Mathematical theory of communication. In: The Mathematical Theory of Communication. Ed. C.E.Shannon and W.Weaver. University of Illinois Press, Urbana, 1949.

[Sharma, 2011] K. L. S. Sharma. Overview of industrial process automation. In: Elsevier, 2011.

[Stutzke, 2001] M. Stutzke and C. Smidts. A stochastic model of fault introduction and removal during software development. Reliability, In: IEEE Transactions on, vol. 50, no. 2, pp. 184{193, 2001.

Authors' Information

***Huiqiang Wang** – The 2nd year PhD student, Institute of Industrial Automation and Software Engineering (IAS), University of Stuttgart, Pfaffenwaldring 47, 70569, Stuttgart, Germany; E-mail: huiqiang.wang@ias.uni-stuttgart.de*

Major Fields of Scientific Research: Reliability and Intelligence , Education on electronics

***Nasser Jazdi** – Scientific staff member, Institute of Industrial Automation and Software Engineering (IAS), University of Stuttgart, Pfaffenwaldring 47, 70569, Stuttgart, Germany; E-mail: nasser.jazdi@ias.uni-stuttgart.de*

Major Fields of Scientific Research: Software Reliability, learning aptitude for industrial automation, Soft Computing for Industrial Automation

***Peter Goehner** – Director of the Institute of Industrial Automation and Software Engineering (IAS), University of Stuttgart, Pfaffenwaldring 47, 70569, Stuttgart, Germany; E-mail: peter.goehner@ias.uni-stuttgart.de*

Major Fields of Scientific Research: agent-oriented concepts for the industrial automation, user-oriented automation, energy optimization of technical systems, learning ability and reliability of automated systems, and concepts for the reuse in the industrial automation

BUSINESS PROCESSES MODELLING WITH DSM PLATFORM AT INTEGRATED SYSTEMS DEVELOPMENT

Anna Lubyagina, Lyudmila Lyadova, Alexander Sukhov

Abstract: *The first and most labor-consuming stage of information systems development is an analysis stage. At this stage the set of different formal models describing systems domains, different aspects of the created system functioning is created. The model-centric approach focuses attention on the models at the each stage of the development process. Modelling tools are used by developers and experts in specific domains. These tools must be affordable for different specialists. Created models must be clarity and understandable for all developers and domain experts. This feature is supported with domain-specific modelling tools. Language workbenches include means for domain-specific languages creation. Modelling with DSM platform is more suitable for the domain experts, system and business analysts. Different tasks of the system design and development request specific formalizing means needed for modeling. Different modelling languages and tools are used by analysts. They need to support continuity of development stages, reusability of created models for the solution of different tasks. DSM platforms provide model transformations, translation of models from one modelling language to another. These tasks are important for large projects of complex information systems development. The basis of language toolkits creation is metamodeling. The tools of MetaLanguage DSM platform, allowing designing domain specific languages and models for different domains, to transform models at systems integration, are described. Examples of metamodels and models, transformation rules are presented.*

Keywords: *modeling, domain-specific languages, language workbench, metamodeling, model transformations.*

ACM Classification Keywords: *D.2 Software Engineering: D.2.2 Design Tools and Techniques – Computer–aided software engineering (CASE); D.2.6 Programming Environments – Domain-specific architectures; Languages (e.g., description, interconnection, definition); D.2.13 Reusable Software – Domain engineering; Reuse models.*

Introduction

Currently, the problem of expediting the development of new informational and analytical systems and configuration of existing systems to meet new challenges, adapt to changing operating conditions and user needs, becomes more urgent. These tasks can only be resolved through the use of models that represent the formal descriptions of domains and requirements to information systems. The use of models, not only during domain analysis and formalization of requirements for information system, but throughout their entire life cycle let to create flexible systems with the maximum capacity of their adaptation. In this regard, tools that are based on the use of model-based approach to create an informational system become more popular. These tools allow reducing the complexity of creating and maintaining systems at the expense of «reusability» of the previously described models to generate and configure the software of information systems.

Model-driven approach to software development involves the use of modeling languages, by which the models are built. The *visual languages* are more commonly used as graphical models have greater clarity and understandable not only for programmers, but also for experts in the relevant domains and end-users.

At designing informational and analytical systems the selected CASE-tools and BI-platforms necessitate the use of appropriate modeling languages. When integrating systems, designed by different developers, it is necessary to revise the models, described with using different languages, execute their *transformation* to allow their integration when creating models of the integrated system.

Another task of the system development efficiency increase is to ensure participation of experts in various subject areas in the process of system development. This requires the creation of appropriate conditions: developed models must be clear for not only to specialists in the field of information technologies, system analysts and programmers, but also for domain experts, who need to work with descriptions, in which the common for these concepts, terminology relevant subject area are used. This is possible if the models are created in terms of domain-specific languages (DSL).

For the development of domain-specific languages and model transformations, special software (language tool-kit, language workbenches or DSM platforms) are used.

Modelling Languages and DSM platforms

During the development of complex systems a lot of models are created. These models are used for solving various tasks, for describing the target system from different points of view, at different levels of abstraction. Thus, a hierarchy of models is created [Lyadova, 2008], in which the models are described by the means of modeling languages (modelling language is used as metamodel), and modeling languages are described by the means of meta-language.

There are many model-based approach implementations that use general-purpose modeling languages for model creation. For example, a general-purpose modeling language UML (Unified Modeling Language), together with the standard MOF (Meta-Object Facility) forms the basis for the concept of MDA (Model-Driven Architecture). But general-purpose modeling languages are often difficult to understand, not only for experts in a particular domain, which are involved in the system creation, but also, in some cases, even for professional developers. In addition, it is sometimes difficult to adequately express the domain concepts, operated by informational system users, using a general-purpose language. That is why the model-oriented software development is increasingly being applied visual domain-specific modeling languages (Domain-Specific Modeling Languages – DSML, Domain-specific Languages – DSL), designed to solve a specific class of problems in a particular domain. DSL are easy to use and understandable to the various categories of professionals, as they operate with usual terms for their subject area. Using a DSL, even users are able to cope with models modification. They can be active participants of the models creation and modification. Users become actors in the informational system development and maintenance, system customizing.

To support the creation of DSLs, a special type of software, called language toolkits, language workbenches or DSM platform (DSM – Domain Specific Modeling) is used. There are various means to create visual DSLs with the ability to define graphical notations: MetaEdit+ ([Kelly, 2013; MetaCase, 2014]) Microsoft DSL Tools ([Cook, 2007; Overview, 2014]), Eclipse GMF ([Sorokin, 2010; Gronback, 2009]), QReal ([Terekhov, 2013; QReal, 2014]), and others. DSM-platforms allow to develop languages and models, as well as make their transformations associated with the need to transition models from one to the other notations, – *model transformation* (translation models, described using one language, into another language).

At the modelling vertical and horizontal transformations are realized. *Vertical transformations* control the process of model transformation during the transition from one level of the hierarchy to the other, such as mapping the metamodel elements to model elements. *Horizontal transformations* are transformations in which the source and target model belong to the same level of the hierarchy. An example of a horizontal transformation is the transformation of a model created by using a one modeling language, to the model described in another language. Means of model transformations can reduce the complexity of informational systems development, providing, on the one hand, the ability to automatically generate data structures and program code, but on the other hand – the ability to export and import models.

There are different approaches to the transformation of models: for example, the system AGG, GReAT, VIATRA uses rewriting graph rules (graph rewriting) to perform the transformations; MTBE approach is based on the programming method for sample [Suhov, 2012]. All currently implemented approaches have various restrictions, complicating their use to describe the transformation (in particular, the need for high skill users). Thus, one of the main problems to be solved when creating language toolkits is the task of developing the means of transformation, removing these restrictions. The implementation of such tools greatly enhances the use of DSM platforms, making them more accessible for different categories of professionals in the development of systems for different purposes and as a basis for the integration of information and analytical systems [Zamyatina, 2013].

Modelling with MetaLanguage DSM platform

System MetaLanguage is a DSM platform, *language workbenches* designed to create a visual domain-specific languages and models with their use, as well as performing the transformations of the created models.

The process of modeling in MetaLanguage includes *several steps* [Suhov, 2013]:

- *Domain-specific language development* (metamodel design) on the basis of metalanguage by using a graphical model editor;
- *Model development* with using the designed language and model validation.

At the *metamodel creation* it primarily needs to determine the basic design of a new language: metamodel entities, the relationship between them, the restrictions on the entities and relationships. The metamodel developer gets a scalable, dynamically configurable visual modelling language as a result of a metamodel constructing. At the process of design of the first level DSL in MetaLanguage the basic constructions of system metalanguage are used: entity, relationship, constraint. In this case, metamodels developed in MetaLanguage can be used as a metalanguage for creating new languages. The process of modelling designing can be iterated. Thus, the hierarchy of models is built.

Using a developed DSL the system analysts can *create models* that contain objects describing the specific entities of the domain and the relationships between them. Developing the model, modelers need to check whether it satisfies the restrictions that have been imposed on model entities and relations. Validation of the created model performs.

When changes are made in the metamodel (in the description of the DSL), the system automatically makes the necessary changes to the model, created by using this metamodel (domain-specific language).

The component of transformation is used for automating model transformations in MetaLanguage.

Vertical transformation is a transformation of the model described in one level of the hierarchy to the model in a different level. Model transformation of the low-level model to the higher-level model of designed hierarchy in MetaLanguage corresponds to the operation of model creating. Reverse transformation allows the interpretation of the higher-level model, determination of the types of model's elements, the implementation of various operations of this model. Algorithms for performing the corresponding operations can automatically maintain consistency of models at different levels.

In order to perform the *horizontal model transformation*, it is needed to set the appropriate transformation rules, which are described using the corresponding metamodels. Rules define the correspondence between the constructions of languages (between the elements of metamodels). The transformation description must be carried out with the use of visual languages understood for the different categories of professionals. In MetaLanguage there is the ability to set horizontal transformations of the «model-to-model» type and «model-to-text» type.

In MetaLanguage system model transformation is based on an algebraic approach with a single ejection based on graph grammars. This approach is suggested in the implementations of a number of modifications ([Lyadova, 2012], [Suhov, 2012]). The transformation component provides the possibility to describe the transformation of element attributes of the metamodel, a multi-level description of metamodels of the left and right side of the production rule. It allows describing the transformation rules on the one hierarchical level, and their application – on the other.

Designing Complex System: Tasks and Approaches to Integration of Heterogeneous Informational and Analytical Subsystems

Now much attention is paid for methods and means of information systems integration, allowing combining a number of different subsystems, to expand their functionality, to improve the efficiency of their use.

The concept «integration» is the key word in the development of complex informational and analytical systems. Integration is needed to solve many problems at the complex systems development. The problem of informational and analytical systems integration is investigated in some directions:

– Data integration is defined as joint of data flows, actual and historical data, received from heterogeneous sources;
– Business processes integration is considered as assembly of common processes of different enterprises at their collaboration;
– Integration of inherited systems means integration of different heterogeneous analytical and informational systems as subsystems at the development of complex system.

In this paper focus is on the functional integration of inherited systems in complex integrated informational and analytical system. For example, national banks need software, which automates and integrates several monitoring operations, implemented by separate subsystems, to meet the challenges of monitoring of non-financial enterprises and their activities. Now much of this monitoring work is not automated. Thus, there is a need to establish a unified organizational and technical system of creating and maintaining summary information about the objects of monitoring obtained from distributed, separate sources. A comprehensive system should consolidate existing subsystems that will solve the problem of interaction of subsystems, combining their functionality [Lubyagina, 2014].

A promising approach to the functional integration of systems is the model-driven design (MDD, Model-Driven Design) and engineering (MDE, Model-Driven Engineering). The systematic use of models throughout the life cycle of informational system is suggested: the definition of data structures, code generation for the user interface and main functionality are model-based. The process of creating software is represented as a chain of transformations of the original model into the ready-to-use information system. In addition, different specialists, domain experts and developers require presentation of models in different languages; they describe models using different graphical notations, the different types of charts.

At the heart of MDE is the concept of DSM (Domain Specific Modeling). At developing the model of comprehensive informational and analytical system based on the existing subsystems the problem of integration of subsystems is solved as integration of models developed for different domains with using different visual languages.

To reduce the labor intensity it is suggested to use the DSM platform that allows reusing designed models.

Approach approbation is applied in the design of complex automation system for monitoring activity of non-financial enterprises with integrating some existing informational and analytical systems.

Design and Transformation of Models in the Integrated Informational and Analytical System with Use of MetaLanguage DSM platform

The basis for the development of domain-specific modeling language for monitoring service activity is a general-purpose (universal) language for modeling in the study domain. This language, on the one hand, must have sufficient expressive power to describe the various activities, automated legacy systems, as well as for modeling complex integrated system. On the other hand, this language must provide the ability to create on its basis new specific languages for modeling various subsystems that take into account their specificities.

The project objectives are:

– To design domain metamodel on the basis of information on the enterprises activities under monitoring (general-purpose modelling language – the basic level of the system modelling);
– To create languages for modelling enterprises in various industries on the basis of the designed general-purpose language metamodel;
– To develop models representing specific business processes of enterprises in the studied domain with use domain specific modelling languages;
– To define transformation rules and execute transformations of developed models to design models of complex system integrating different models described on DSLs.

Figure 1 shows a simplified general-purpose language metamodel created with using MetaLanguage system.

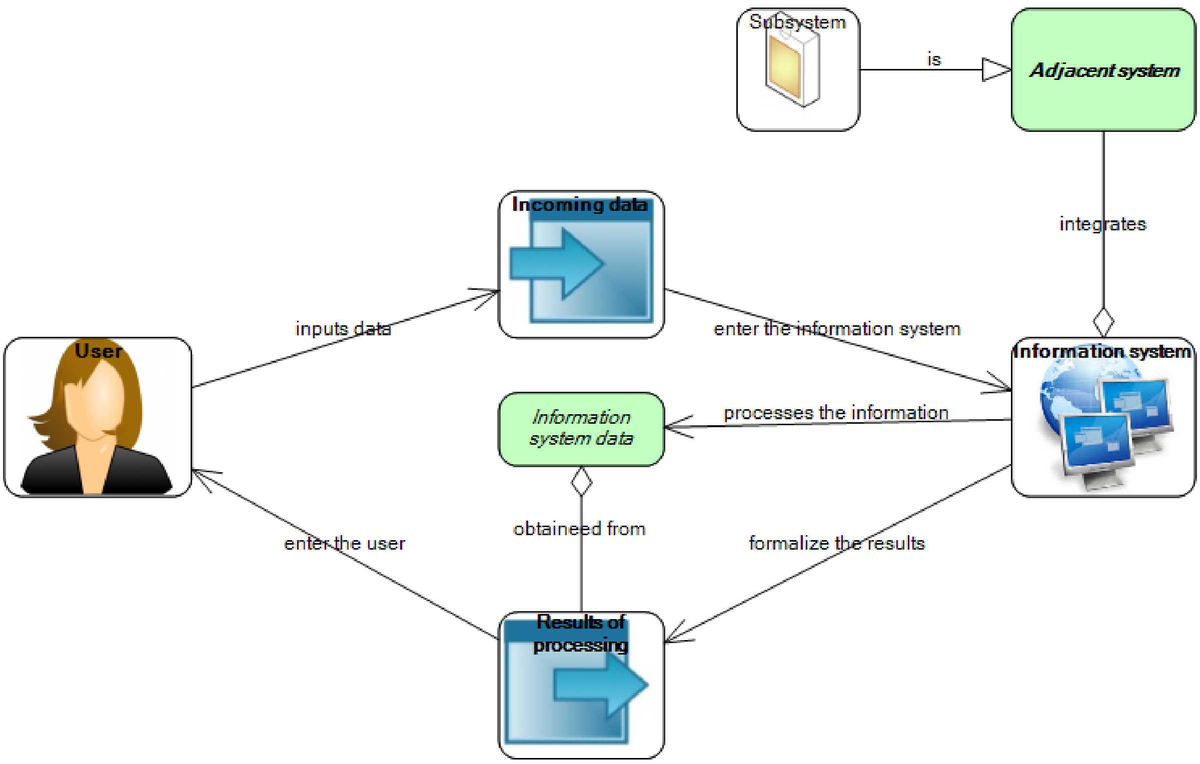


Figure 1. Simplified general-purpose language metamodel built with using the MetaLanguage system

Created general-purpose DSL has the following advantages: a sufficient expressive power, ensuring adequate representation of business processes in the study area; metamodel for ease of understanding, the high degree of its transparency; focus on the tasks to be solved in the development of a comprehensive monitoring system that allows users to easily learn the language and work with models using familiar engineering terminology.

In implementing the metamodel language were created following entities (see Fig. 1):

- «*Subsystem*» – this entity represents subsystems that are the parts of an integrated information system;
- «*Adjacent system*» – this entity represents any subsystem of the integrated system;
- «*Information system*» – this entity represents the complex (integrating) system;
- «*User*» – this entity represents all user roles for users that have access to work with an integrated system;
- «*Incoming data*» – this entity represents input data, incoming documents of the complex system;
- «*Information system data*» – this entity represents interim data at various stages of processing;
- «*Results of processing*» – this entity represents output data, reports in the integrated system.

Entity «*Subsystem*» has a relationship of *inheritance* type with «*Adjacent system*» («*is*»). Entities «*Adjacent system*» and «*Information system*» are connected by relationship of *aggregation* type («*integrates*»). The remaining relationships of entities are type of *association*.

A model of application execution (bid processing) by members of the monitoring group is designed on the base of general-purpose modeling language. Figure 2 shows a simplified version of this model.

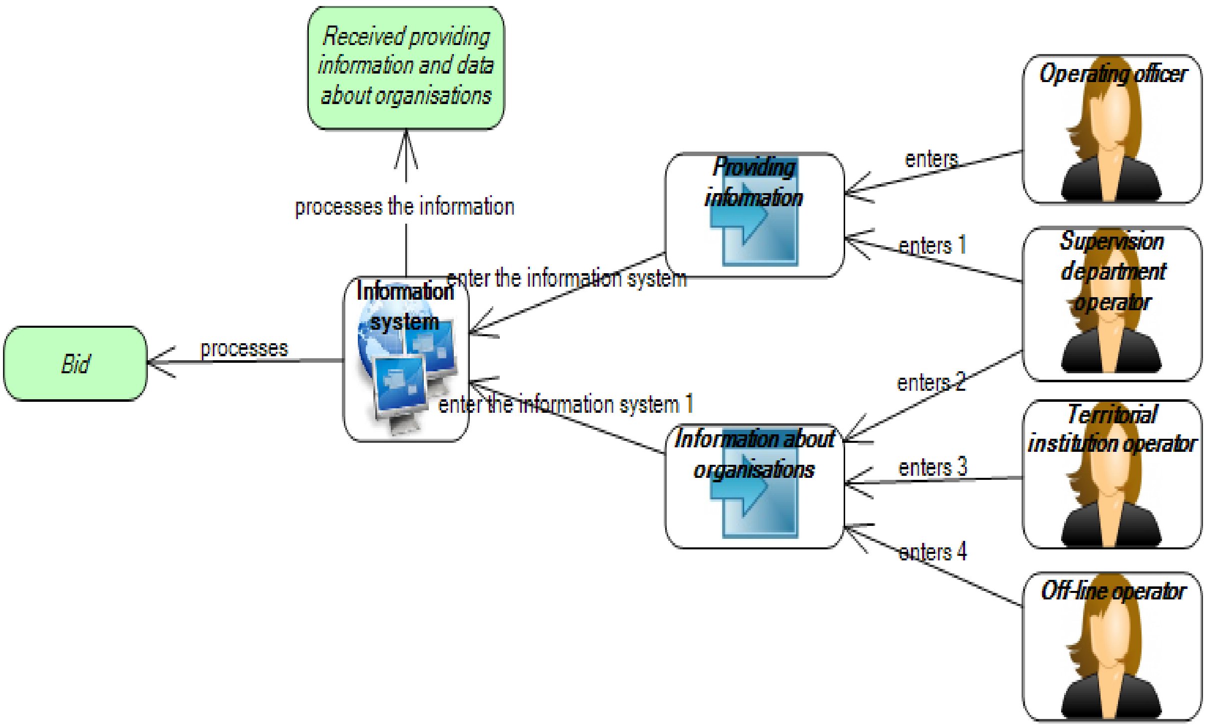


Figure 2. «Bid processing» model, built on the basis of the general-purpose metamodel

The model constructed with using the general-purpose language cannot represent all the operations of the application processing of monitoring system in detail. In addition, the constructed model does not reflect the partition of an integrated system for specific subsystems; functionality of each subsystem and the extent of their use in the treatment are impossible to distinguish. Thus, the constructed model has a number of shortcomings: lack of visibility and specific details for end users, the unsufficient reflection of the specific functionality of each subsystem.

Modelers need to establish specific languages for specific tasks of various subsystems for specific categories of users in order to overcome these shortcomings.

The general-purpose language (metamodel is shown above) can be used as a metalanguage for designing the second level DSLs. These languages allow describing more clearly the business processes of systems that automate the different tasks of monitoring. New object-oriented language is created in a specific syntax notation of the general-purpose modeling language (see Figure 3).

More detailed models are built on the basis of the new constructed object-oriented metamodel. An example of the simplified model of application processing by members of the monitoring group is shown below (see Figure 4).

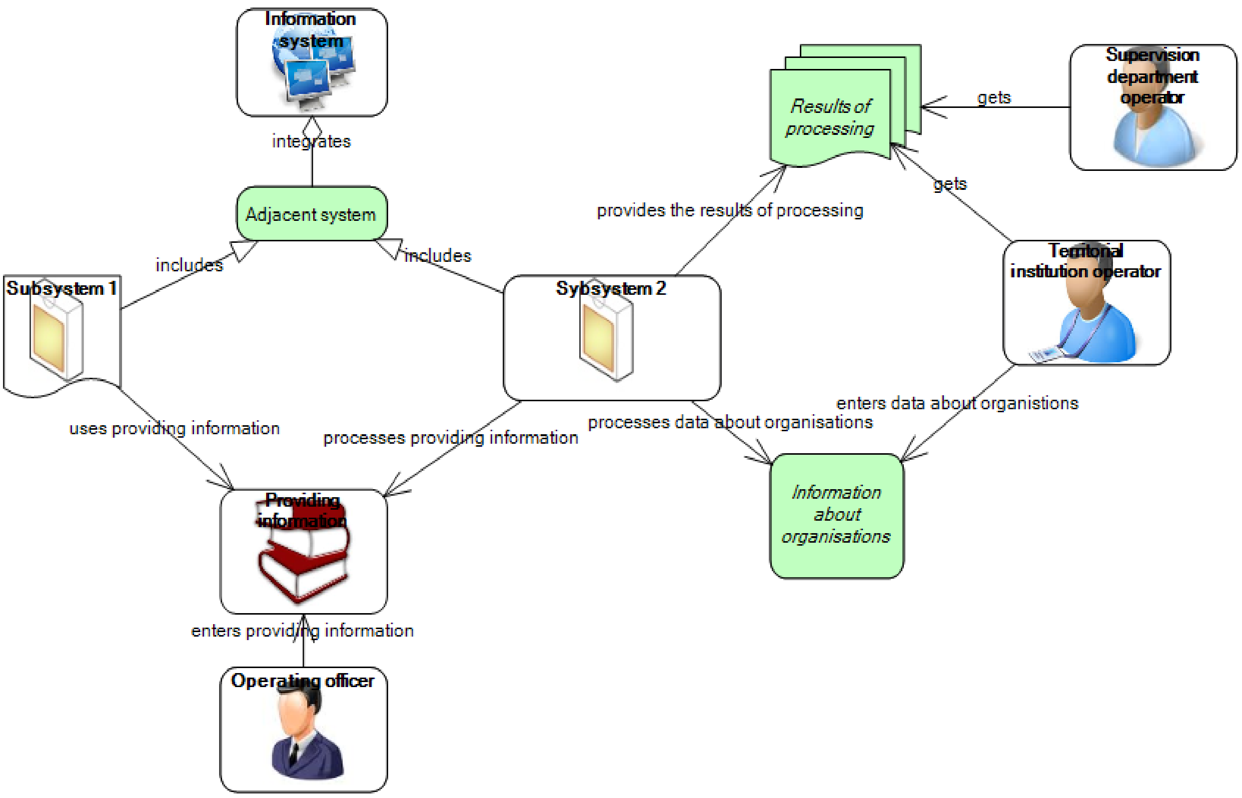


Figure 3. A simplified metamodel, built on the DSL of the first level

This model (Figure 4) is developed with using specific object-oriented language of the second level. It has more detail, reflects the specific end-user experience and functionality of the subsystem, and allows working in terms familiar for users of subsystem.

Modeling languages to describe the other subsystems can also be created on the basis of a general-purpose modeling language.

These examples demonstrate the possibility of the suggested approach to ensure the creation of models that can be set to the desired level of detail and reflect the specificity of the subsystems and users. Modelling languages and models for the two pilot subsystems of integrated informational and analytical system are developed at the research project execution.

However, to meet the challenges of the project, there was a need to translate the constructed models into other languages to transfer the results of the modelling to developers of the complex informational system. In particular, developers need to build ER-diagrams, Use-Case diagrams, etc. Automated translation of the constructed models from designed domain specific notations into standard notations needed for developers were executed with the transformation tools of MetaLanguage.

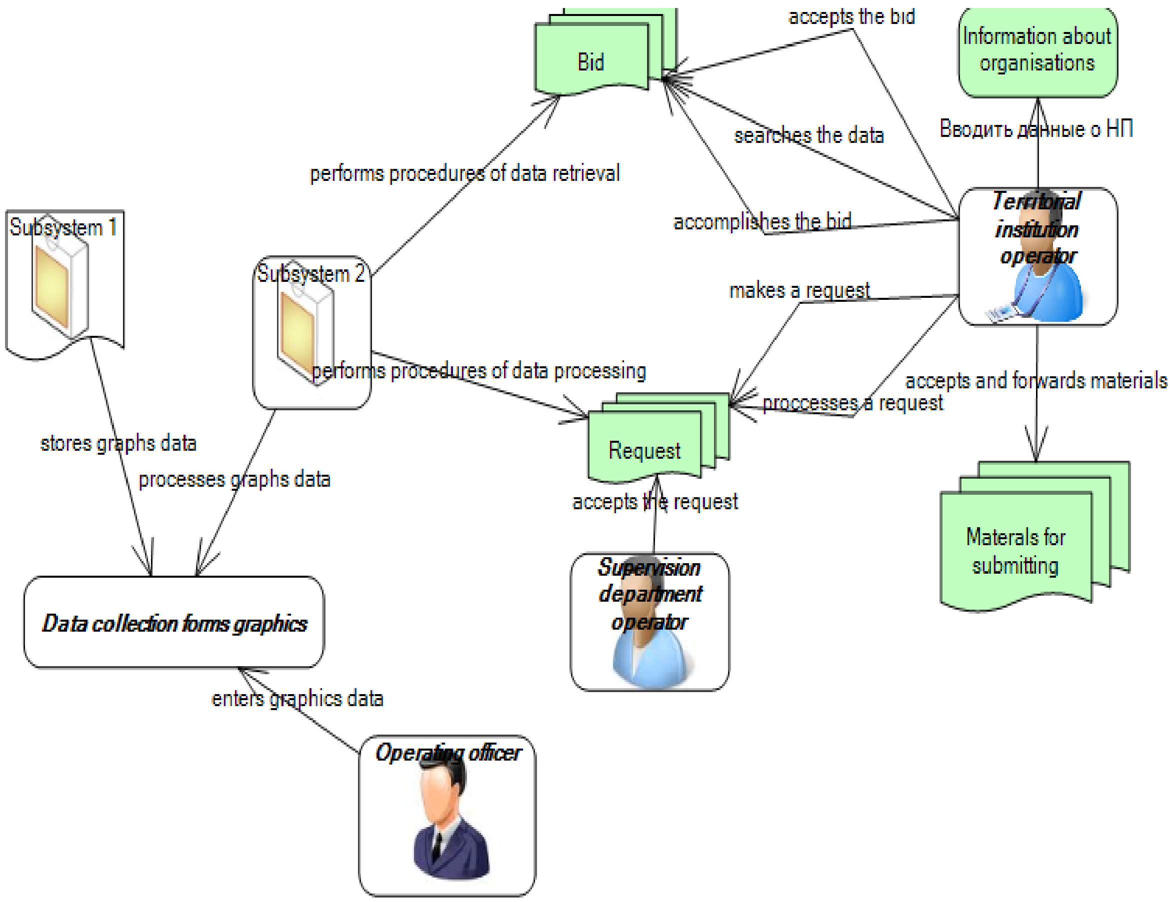


Figure 4. A simplified «Application processing» model, built on the DSL of the second level

To translate models by using the transformation component of MetaLanguage, developers have to execute two steps:

– To describe transformation rules using metamodels of the source and target models (languages);
– To make a model transformation (to translate the created models into the target language in accordance with the described rules).

Transformation rules are described by using visual constructor: developer has to define the left and right parts of the rule. Constructions of the source and target languages are used in the rule definition. If the transformation rules are given, then the second step is performed automatically in accordance with these rules. Moreover, the described rules can be applied to the transformation of models that can be described by using the same languages later, significantly reducing complexity of the translation. A set of rules can be expanded later. In the description of the rules, as well as descriptions of modelling languages, it is possible to make changes.

To translate the models described on the created DSL into the notation of UML (activity diagrams), metamodel shown in Figure 5 was used.

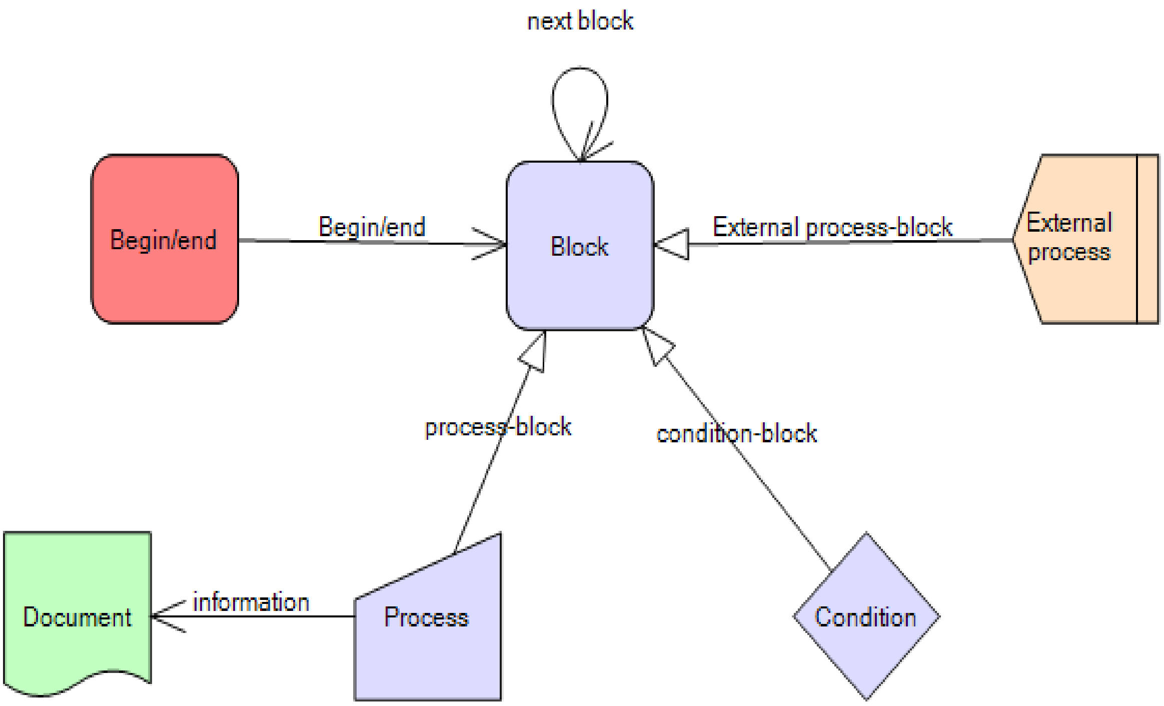
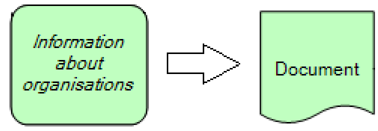
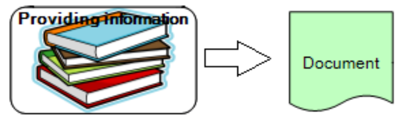
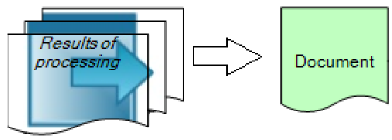
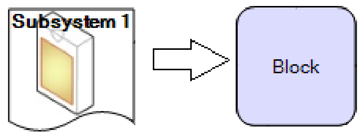
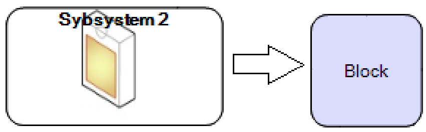
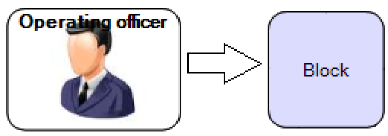
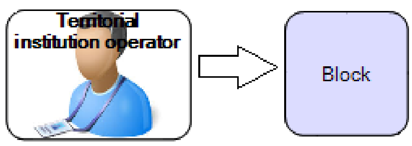
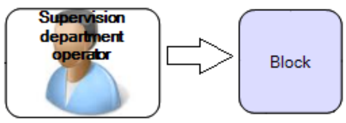
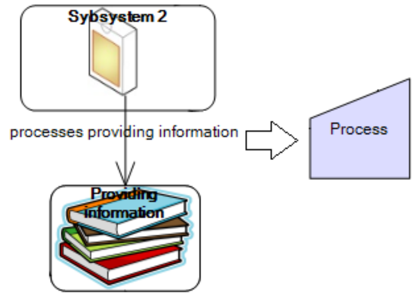
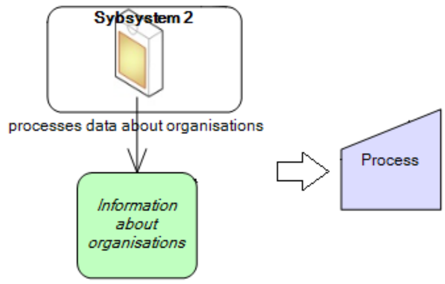
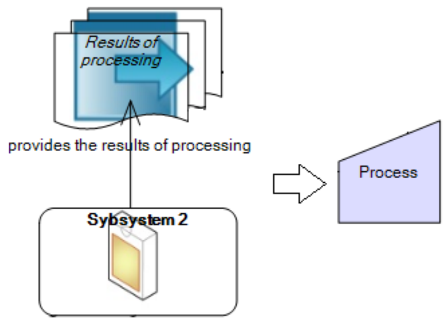
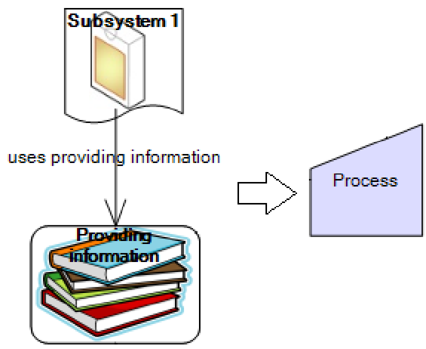


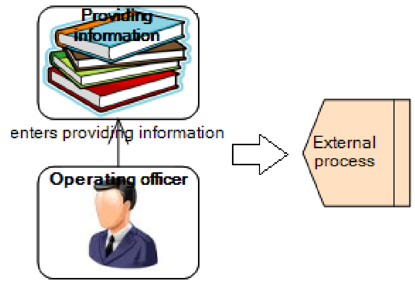
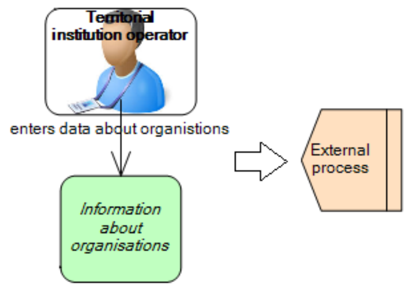
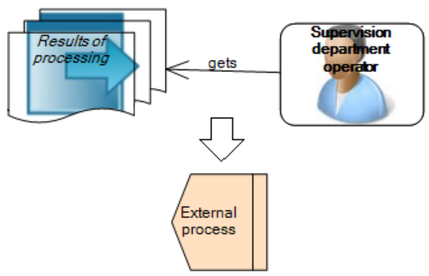
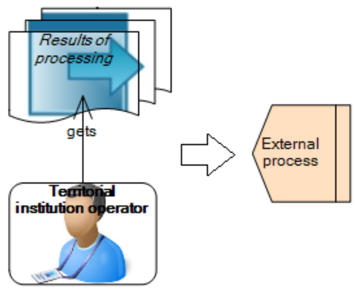
Figure 5. Metamodel of activity diagram in MetaLanguage system

Transformation rules developed in the MetaLanguage system for translation of the constructed models for language UML (for building activity diagrams) are shown in Table 1. Figure 6 shows the result of the transformation of the «*model-to-model*» for the simplified model «*Application processing*».

Table 1. Transformation rules of «application process» model in the UML notation (activity diagram)

Transformation rule (left and right parts)	**Graphical mapping**
Information about organizations – Document	
Providing information – Document	
Result of processing – Document	
Subsystem 1 – Block	
Subsystem 2 – Block	
Operating officer – Block	
Territorial institution operator – Block	
Supervision department – Block	

Transformation rule (left and right parts)	Graphical mapping
Processes providing information – Process	
Processes data about organizations – Process	
Provides the results of processing – Process	
Uses providing information – Process	

Transformation rule (left and right parts)	Graphical mapping
Enters providing information – External process	
Territorial institution operator – External process	
Gets – External process	
Gets – External process	

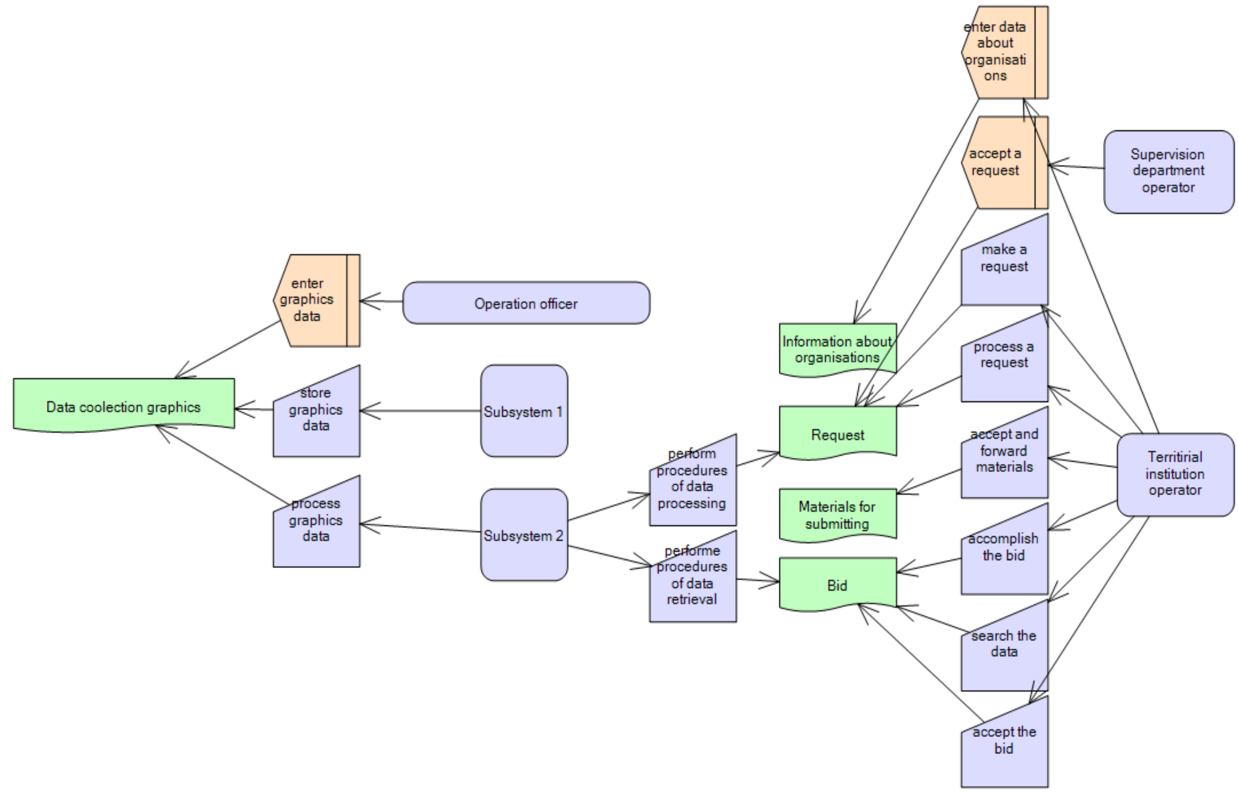
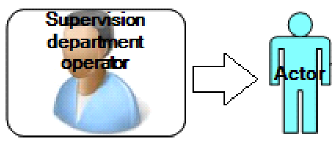
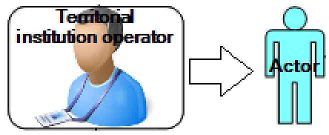
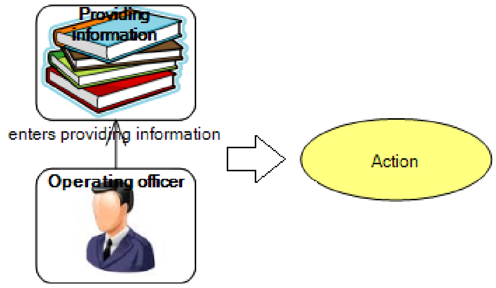
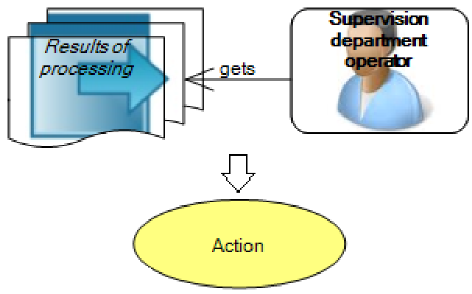
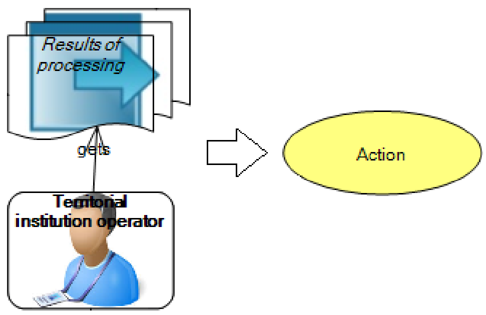
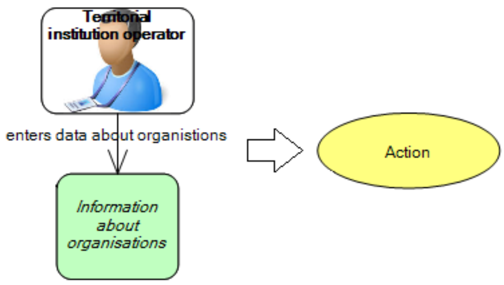


Figure 6. The result of the model «Application Processing» transformation – Activity diagram

In addition, the model of the application processing has been transformed into Use-Case diagram notation. Figure 7 presents a metamodel of Use-Case diagram, and Figure 8 shows the result of transformation of the model formed in accordance with the rules shown in Table 2.

Table 2. Transformation rules of «Application processing» model into the UML notation (Use Cases Diagram)

Transformation rule (left and right parts)	Graphical mapping
Operating officer – Actor	
Supervision department operator – Actor	

Transformation rule (left and right parts)	Graphical mapping
Territorial institution operator – Actor	
Enters providing information – Action	
Gets – Action	
Gets – Action	
Enters data about organizations – Action	

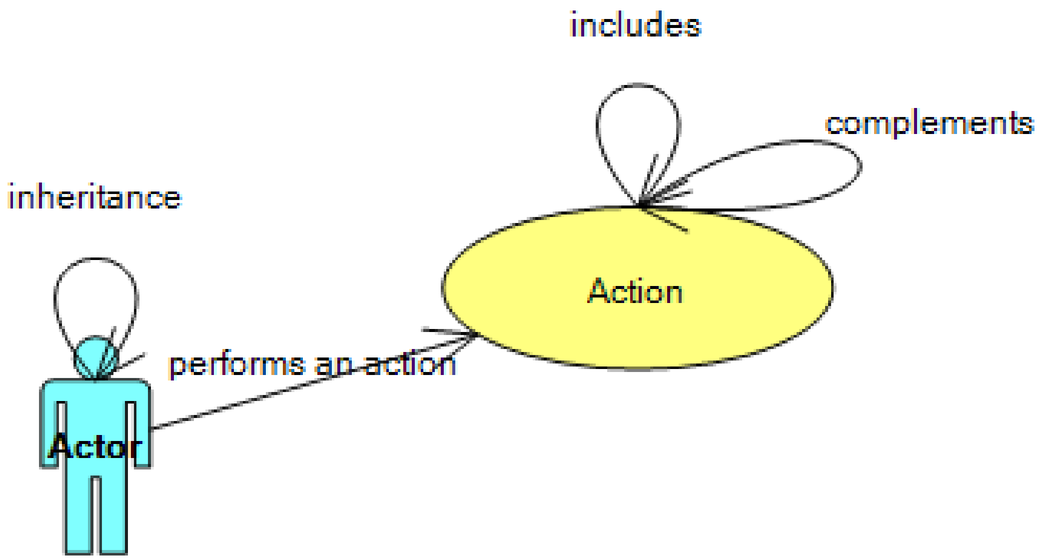


Figure 7. Metamodel of Use Cases Diagram

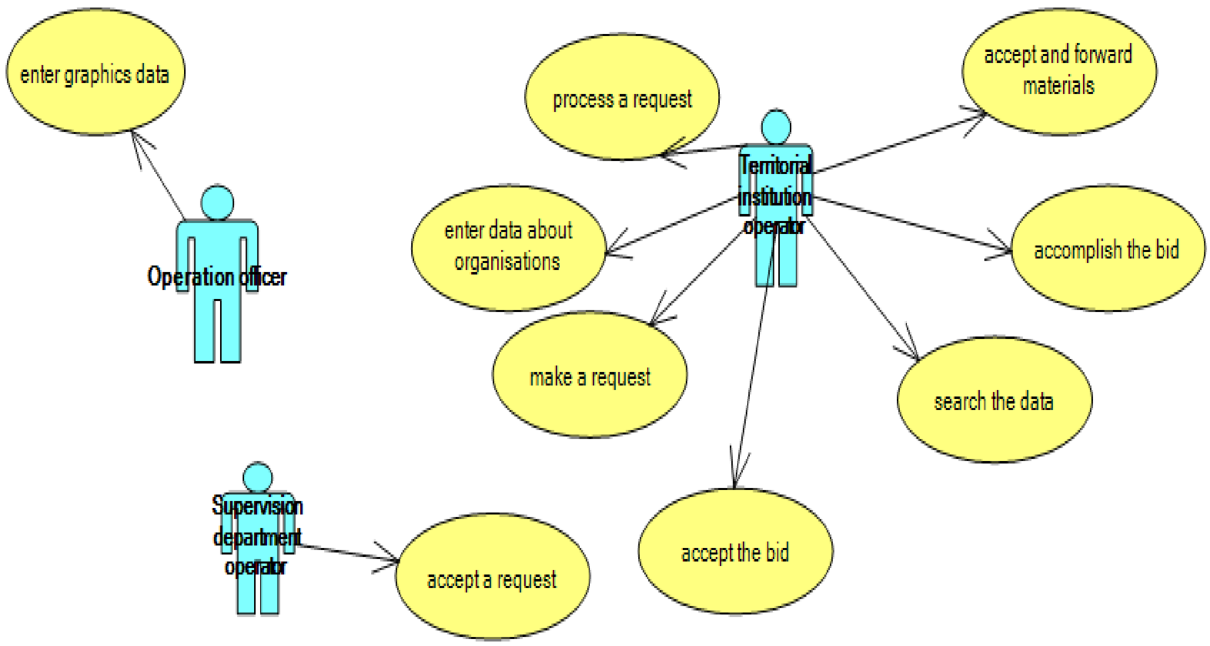


Figure 8. The result of the model «Application Processing» transformation – Use-Case diagram

Description of «model-to-model» transformation does not require any knowledge of any specific programming language. All operations are implemented in the same environment and at the same level of skill requirements of users as to the development of languages and models.

At the research project also were described the rules and made the transformation of the type «model-to-text» to generate code in the language of SQL. The resulting code can be used for creation of the integrated information system database.

Conclusion

The research project has shown that the modern DSM platforms can be used not only for the development of domain-specific languages and models, but also as a means of integration of various informational and analytical systems. Earlier domain-specific modelling languages were developed for creation of simulation models, transformation rules were described and transformations of the developed models for simulation experiments in GPSS were executed.

The ability to dynamically change of languages and models, their configuration in the development of informational and analytical systems can significantly reduce the complexity of problems solving of modeling due to re-using previously established models, automation of their transformations.

In addition, the modeling tools, based on using language toolkits that provide advanced tools for the development of languages and models, do not need high requirements for the qualification of users; these workbenches are available for different categories of users.

Bibliography

[Cook, 2007] S. Cook, G. Jones, S. Kent, A.C. Wills. Domain-specific Development with Visual Studio DSL Tools. Reading: Addison-Wesley, 2007. – 560 p.

[Gronback, 2009] R.C. Gronback. EclISse Modeling Project: A Domain-specific Language Toolkit. In: Reading: Addison–Wesley, 2009. – 706 p.

[Kelly, 2013] S. Kelly, K. Lyytinen, M. Rossi, J.P. Tolvanen. MetaEdit+ at the Age of 20. In: Seminal Contributions to Information Systems Engineering. Springer, 2013. P. 131-137.

[Lubyagina, 2014] A.O. Lubyagina. An approach to the functional integration of technology-based MDE. In: Izvestiya of SFU. Engineering science. – 2014. – № 6 (155). P. 159-163.

[Lyadova, 2008] L.N. Lyadova. Multilevel models and languages DSL as a basis to create intelligent CASE–systems. In: Proceedings of international scientific and technical conference «Intelligent systems» (AIS'08) and «Intelligent CAD» (CAD–2008). A scholarly edition in 4 volumes. T. 2. Moscow: Fizmatlit, 2008. P. 37- 41.

[Lyadova, 2012] L.N. Lyadova, A.P. Seriy, A.O. Sukhov. Approaches to the description of the vertical and horizontal transformations metamodels. In: Mathematics of Program Systems: Collection of scientific papers. – Perm State University, 2012. – Vol. 9. – P. 33-49.

[MetaCase, 2014] MetaEdit+ Version 5.0: Workbench User's Guide. [http://metacase.com/support/50/manuals/mwb/Mw.html] (Checked: 18.04.2014).

[Overview, 2014] Overview of Domain-specific Language Tools. [http://msdn.microsoft.com/en–us/library/bb126327.aspx] (Checked: 18.04.2014).

[QReal, 2014] QReal: BP. [http://qreal.ru/static.php?link=QRealBP_usage] (Checked: 18.04.2014).

[Sorokin, 2010] A.V. Sorokin, D.V. Koznov. Overview of EclISse Modeling Project. In: System Programming. MY. 5 S. 2010.

[Suhov, 2012] A.O. Sukhov. Methods for transforming visual models. In: Proceedings of the III International Scientific and Technical Conference «Technologies of Information Systems Development» (TRIS–2012). Taganrog Technological Inst. of SFU, 2012. – Vol. 1. – P. 120-124.

[Suhov, 2012] A.O. Sukhov, A.P. Seriy. Using graph grammars for model transformation. In: Proceedings of the conference «CSEDays 2012». Ekaterinburg: Ural State University, 2012. – P. 48-55.

[Suhov, 2013] A.O. Sukhov. Tools for creating visual Domain-specific modeling languages. In: Fundamental research. – 2013. – № 4. – P. 848-852.

[Terekhov, 2013] A.N. Terekhov, T.A. Bryksin, Y. Litvinov. QReal: platform visual object-oriented modeling. In: Software Engineering, 2013, № 6. P. 11-19.

[Zamyatina, 2013] E.B. Zamyatina, L.N. Lyadova, A.O. Sukhov. An approach to the integration of simulation and information systems on the basis of DSM-platform MetaLanguage. In: Proceedings of the IV International Scientific and Technical Conference«Technologies of Information Systems Development» (TRIS–2013). Taganrog Technological Inst. of SFU, 2013. – Vol. 1. – P. 61-70.

Authors' Information

Anna Lubyagina *– National Research University Higher School of Economics, Department of Business Informatics, student; 38, Studenceskaia St., Perm, 614070, Russia; e-mail: lubyagina@prognoz.ru.*

Major Fields of Scientific Research: Business informatics, Business process modelling, Analytical systems, System analysis.

Lyudmila Lyadova *– National Research University Higher School of Economics, Department of Business Informatics, associate professor; 38, Studenceskaia St., Perm, 614070, Russia; e-mail: LNLyadova@gmail.com, LLyadova@hse.ru.*

Major Fields of Scientific Research: Information systems development, Modelling languages, Domain specific modelling, CASE-tools, Graph models.

Sukhov Alexander *– National Research University Higher School of Economics, Department of Business Informatics, teacher; 38, Studenceskaia St., Perm, 614070, Russia; e-mail: ASuhov@hse.ru.*

Major Fields of Scientific Research: Software Engineering, Modelling languages, Visual modelling, Domain specific modelling, Domain specific language, Language workbenches

IMPROVING AUTOMATIC SPEECH RECOGNITION ACCURACY BY MEANS OF PRONUNCIATION VARIATION MODELING

Vladimir Chuchupal, Anton Korenchikov

Abstract: *We explore the properties of the pronunciation variation (PV) models as an approach for an automatic speech recognition accuracy improvement. The PV model is formulated as well as the methods to find out PV parameters and include the model into the search procedures. We show that utilizing of the PV models could substantionally increase the accuracy of automatic recognition of natural speech.*

Keywords: *automatic speech recognition, acoustic modeling, speech pronunciation variation modeling, explicit models, hidden markov models.*

ACM Classification Keywords*: I.2.7 Natural Language Processing - Speech recognition and synthesis*

Introduction

The pronunciation of a word in a speech recognition system (ASR) usially is determined by its pronunciation or phoneme transcription. As a rule most of the words has a single pronunciation transcription, namely the basic or canonical one.

In spontaneous speech a pronunciation may substantionally differs from the canonical and this is one of the most important sources of the errors of speech recognizers.

There are currently two approaches to pronunciation variation (PV) modeling for ASR [Wester, 2003; Fosler, 1999]. The explicit modeling describes all probable pronunciation variations in terms of explicit changes of the basic word transcriptions. In the other words, in explicit modeling the given word pronunciation could be defined as a set of the most probable word transcriptions. The implicit modeling [Saraclar, 2004] describes the variations in pronunciation by means of changes in the structure of the allophone hidden Markov models of the basic transcription.

The both approaches do not eliminate the need to use the basic transcriptions.

The correct implementation of pronunciation variation models may have a great impact on the accuracy of ASR. Such a conclusion is follows from the heuristic analysis of the errors done by ASR as well as the oracle-style experiments. It had been showed that using of the adequate phonemic transcriptions reduce word error rate (WER) as much as nearly a twice [Saraclar at al., 2000].

At the same time the reported improvements in WER obtained with the use of pronunciation variation models still are far from the expected ones. For example on Dutch corpora VIOS the WER decrease obtained was about 0.8% (from initial level of 10.7% to 9.9%) at the significant number of the pronunciation variants: around 4.9 per a vocabulary word [Wester, 2003]. On Switchboard corpora the implementation of the implicit pronunciation models led to decrease in WER at 1.7% (from 39.4% to 37.7%) [Saraclar, 2004]. On NIST–2000 Hub-5 data the use of pronunciaton variation models improved WER at 2.2%: from 54.6% down to 52.4% [Zheng, 2003].

In this study we implemented the pronunciation variaton model in existing Russian ASR. We follow the explicit approach to pronunciation variation modeling in that all changes in pronunciation could be adequately described in terms of deletions, substitutions and insertions of the phonemes.

For implementation of this approach we have to address the following issues:

- Define a pronunciation variation models,
- Find the most probable phone transcriptions for the words,
- Find a method for estimating the parameters of the pronunciation model,
- Find a use of implementation of pronunciation models in the search procedure.

Pronunciation variation model

$$W = \underset{W}{arg\,max}\, P(W|X) = \underset{W}{arg\,max}\, \frac{P(X|W)P(W)}{P(X)}. \quad (1)$$

Note that the important distinction between words and phone transcriptions is that a word defines or relates to the meaning whereas a transcription defines the acoustic image of the word. Such a difference could be accounted in the framework of known statistical approach to speech recognition [Bahl, 1983].

Let $X = \{x_t\}, t = 1, \ldots, T$ be a sequence of the vector parameters of the observed speech signal, $W = \{w_i\}, i = 1, \ldots, N$ be a sequence of the vocabulary words. The result of recognition of X, the most probable word sequence W^*, can be obtained from the equation [Jelinek, 1997]

$$W^* = \underset{W}{arg\,max}\, P(W|X) = \underset{W}{arg\,max}\, \frac{P(X|W)P(W)}{P(X)}. \quad (2)$$

The first factor $P(X|W)$ in the numerator (2) denotes the data likelihood conditioned the given word sequence and could be obtained with the help of the acoustic phone models. The value of second factor $P(W)$ is estimated with the help of the language model.

We use t^w to denote the phonemic transcription (pronounciation model) of a word w. The set of phonemic transcriptions of the given word w is denoted as T^w. The pronuncicaion model for the word sequence W is denoted T^W. The designation t^W will be used as a notation for the arbitarry sequence of word phonemic trancsriptions from T^W.

The conventional speech decoding and recognition procedures as a rule defines the best sequence of acoustical models (phonemic transcriptions), not the best sequence of words, that is instead of (2) de facto is used:

$$t^{W^*} = \underset{t^W}{arg\,max}\, \frac{P(X|t^W)P(t^W)}{P(X)}. \quad (3)$$

Then the most probable word sequence could be obtained by mapping of each pronunciation model to the corresponding word, i.e.:

$$t^{W^*} \rightarrow W^*. \quad (4)$$

If for all words there is a single pronunciation per word in vocabulary the methods (2) and (3) are equivalent.

Using the identity $P(t^W) = P(t^W|W)P(W)$ the expression (3) could be written as:

$$W^* = \underset{t^W}{arg\,max}\, \frac{P(X|t^W)P(t^W|W)P(W)}{P(X)}. \quad (5)$$

The expression (5) differs from the one of (3) in that it contains the factor $P(t^W|W)$ that accounts the pronunciaton variation. The set of probabilities $P(T^W|W) = \{P(t^W|W), t^W \in T^W\}$ is considered as parameters of the PV model.

Estimation of parameters of pronunciation variation modell

In order to utilize (5) we need to know the parameters of the three types of models: acoustic models, language model and the pronunciation model.

The language model parameters for estimation $P(W)$ usially considered as an independent of the acoustic models. Therefore the estimation of language model parameters could be performed in the independent manner exactly as it is done in conventional (9) approach.

The pronunciation model parameters $P(T^W|W)$ are dependent on the acoustic training data, therefore the independent (of acoustic one) estimation of $P(T^W|W)$ is not correct.

Consider the maximum likelihood estimate of the pronunciation model parameters.

Suppose that the traninig corpora X is such that for all its utterances we know the sequence of words $w_1 w_2 \ldots w_N$ as well as a sequence of the phonemic transcriptions $t_1^w t_2^w \ldots t_N^w$. In such a case the most probable estimate of the parameters $p(t^w|w)$ will be obtained by solving:

$$p(t^w|w) = \underset{w,t^w}{arg\,max} \prod_{w,t^w} p(t^w|w). \tag{6}$$

This frequency estimate is similar to the estimate for the n–gram language model[Young, 1997]:

$$p(t^w|w) = \frac{\#\{t^w\}}{\#\{w\}}, \tag{7}$$

where $\#$ denotes the number of events in curly braces, encountered in the traning data. Therefore the most probable estimate for the given transcription will be the relative frequency of that transcription it in the traninig corpora.

Since the independent estimation of the acoustic and pronunciation model parameters is not correct consider the algorithm consisting of two-step iterations.

Suppose that there is a traninig speech corpora along with the vocabulary. Suppose that for each vocabulary word we know all of the pronunciation variants and consider for beginning the variants as equally probable.

On the first step the maximum likelihood estimates of the acoustic model parameters are obtained. The conventional training methods based on forward-backward and Baum-Welch algorithms can be used.

Then make the co-called «restricted» recognition of all utterances in the speech corpora. Term «restricted» means that the purpose is to find out the most probable sequences of phones given the true word sequences.

On the second step using (7) the maximum likelihood estimations of PV model parameters is obtained. It is done with the help of the co-called «restricted» recognition of all utterances in the speech corpora. Term «restricted» means that the true word sequence is known in advance and the target is to find out the most probable sequences of phones or, another words, sequence of transcriptions.

Then repeat the step 1 and retrain the acoustic models using the obtained on the step 2 the most probable sequencies of phones.

There steps can be repeated for the fixed number of times or until some stopping criteria will be reached.

The embedding of the pronunciation variation modeling into speech decoder

A conventional way to use the several pronunciation transcriptions per word in speech decoder consists of inclusion of each transcription to the pronunciaton vocabulary and treating this transcription in an independent manner as if it is a transcription of a new word. This approach implies no changes in the search algorithms (3)-(4).

It is not the optimal solution though.

Rewrite the expression (2):

$$P(W|X) = \frac{P(W,X)}{P(X)} = \frac{\sum_{t^W \in T^W} P(X,t^W)}{P(X)} = \frac{\sum_{t^W \in T^W} P(X|t^W)P(t^W)}{P(X)}. \quad (8)$$

From (5) and (8) it follows that the most probable sequence of words W^* should satisfy

$$W^* = \underset{W}{arg\,max} \sum_{t^W \in T^W} P(t^W|X)P(t^W). \quad (9)$$

Solution (9) let us define the most probable word sequence (not a most probable phone or transcription sequence) that is exactly what we need from the speech recognition system.

The algorithm (5) differs from the one of (3)-(4) in that we need to take into account the relative frequencies of word phone transcriptions and make the final decision using the weighted sum of the transcription likelihoods.

To implement (9) we need to make an additional, as compared to (3)-(4) calculations accordingly to (9).

Since for every word w:

$$P(w) = \sum_{t^w \in T^w} P(t^w|X), \quad (10)$$

then if, for example, a prefix tree lexicon representation is used, in every tree leaf the word likelihood should be estimated accordingly to (10).

Comparing with the conventional (3)-(4) approach it is also necessary to make some trivial changes in the search data structures and the memory allocaiton.

The practical implementation of the (9)-(10) is associated with the difficulty because of tree pruning [Young, 1997]. Suppose that some leafs of the prefix tree have been pruned because of the relatively small likelihoods. In such a case the likelihoods of these leafs are not known and the corresponding transcriptions could not be used in (10).

To overcome that difficulty consider the following version of (10):

$$W^* = \underset{W,t^W}{arg\,max}\, P(t^W|X)P(t^W). \quad (11)$$

Here the weighted sum of the likelihoods is replaced with the likilihood of the most probable transcription penalized with $P(t^W|X)$.

Numerical experiments

The performance of the considered PV models have been compared on the speech corpora ISABASE–2 [Bogdanov, 2004] and TeCoRus [Chuchupal, 2005]. The training data of the first test consisted of speech utterances of 200 speakers of ISABASE–2 (40K utterances) Ðÿ 50 speakers from TeCoRus (3K utterances). The test material consisted of the 776 utterances that contained the connected digit strings (3147 digits). The vocabulary has been limited to the digits. The reason to use numbers was that the numbers and numerals could provide a lot of examples of pronunciation variations.

No language models has been used.

The recognition results in terms of word error rate (WER) values are presented in Tabl. 1. The column «Basic» contains the results for the case when the basic transcription is used only. The column «Convent.» correspondes the method (2 - 4). The column «Optim.» containes the results for the method (9). Column «SubOptim.» contains results for the method (11). The row «Variability» contains the mean number of transcriptions per vocabulary word.

The results depicted above could be interpreted as an evidence of lack of pronunciation variability in the test corpora. It can be true because the speakers of TeCoRus belong to the same high-edicated profeccisonal group and were born and living in Moscow region. The test material contained a read, carefully articulated speech.

Table 1: Word Error Rate for some pronunciaion variation models (TeCoRus data only)

Method	Basic	Convent.	Optim.	SubOptim.
WER	1.62	5.78	2.00	3.17
Variativity	1.0	1.9	1.9	1.9

The lack of the PV in the first test could explain the observed behavior of the training algoritm: on the TeCoRus data with the increasing number of iterations the mean number of transcriptions per word converged to one.

To obtain recognition results for the data with actual pronunciation variabiliry the second recognition experiment had been fullfilled. The training set of the second test was the same as in the first test. The test set consisted of 867 utterances of 11 test TeCoRus speakers. These data mostly consists of the sequences of numbers and numerals. The vocabulary of the test set consisted of 129 words. Test utterances also contain an additive and casual types of office noise as well as amount of the speech disfluencies that often led to the speech recognition errors.

The pronunciation vocabulary contains 129 numerals.

Talbe 2 shows the WER result for the second test. The table column «Convent.» shows the WER value for the case when the basic pronunciations were used only.

Table 2: WER value for some pronunciaion variation models with the TeCoRus extended data

Method	Basic	Convent.	Optim.	SubOptim.
WER	7.78	7.57	7.38	7.44
Variativity	1.0	1.3	1.3	1.3

The results drawn in (2) could be considered as more relevant to the expected. The best approach appears to be the one that corresponds to the frequency weighting of the pronunciation variants(9). The approach with the inclusion the rival transcription to the pronunciation vocabulary (2 - 4) appears to be less effective both the (9) as well as algotithm (11). In all cases the inclusion of the pronunciation variations appears to be more effective than the use of the basic transcriptions only.

The WER improvements in the second test were not so substantial as it could be expected though. On the one hand it could be because of the type of test material. At the same time the WER improvements observed might be because of the speech corpora TeCoRus and Plantronics had been collected in different conditions. TeCoRus had been recorded with the Senheiser professional microphone while ISABASE–2 corpora had been recorded with the cheap Plantronics microphones.

To clarify these issues the third recognition test had been performed on the speech corpora that contained the natural spontaneous speech that had been extracted from the radio interviews. We used the interviews downloaded from the radiostation «Echo Moscow» [Echo Moscow].

The initial set of rival pronunciation transcriptions for the numerals as well as their relative frequency were the same as in the previous test.

The inteviews have been automatically segmented. The utterances with the numerals have been found and extracted to the separate speech files. The test set consisted of 200 speech utterances of 2–4 words each, with total vocabulary of 91 words.

No language models were used during recognition.

Table 3 presents the results for this test. The table column «Equal.» contains the WER values for the method (9) in the case when the equal relative frequencies for all rival transcriptions were used.

The substantionally higher WER values obtained because of the lack of language model, mismatch between training and testing conditions for acoustic models, and noisy environment during an interviews.

Table 3: WER values for the differnet types of pronunciation variation models on the natural fluent spoken speech

Method	Basic	Convent.	Optimal.	SubOpt.	Equal
WER	69.3	57.44	59.7	60.0	59.5

In third test the observed relative improvements in WER was from 13,4% to more than 17,1% comparing to 5% relative improvement in previous test.

It is shown therefore that for fluently spoken numerals the use of PV models can lead to the substantional improve the accuracy of speech recognition.

Note that there are the other (besides of pronunciation changes) possible reasons of improvements the accuracy of recognition in third test are exists. There is a significant mismatch in the traning and testing data. The test data coded in MPEGt. Howeve if it was the case then the similar WER improvements were to take place in the second test. It had not happened though.

The observed absence of improvement in WER (compared with the other methods) for the methods with weighting of rival transcriptions can be explained from the point of the language modeling. The transcription weighting as well as using the number of rival word transcriptions for numerals has an effect that is similar the using of the unigram languge model. In the test material the relative numeral frequencies were much higher than for the other. The use of conventional method has an effect of using the bigger unigram weights for numerals that was relevant to the data of the test corpora.

Conclusion

The research of the methods for improving the automatic speech recognition accuracy through the use of pronunciation variation models is fulfilled. The probabilistic pronunciation variation model is formulated and well as the ways to estimate the model parameters. The numerical experiments shows that the implementation of the pronunciation variation models is an effective way to improve an accuracy of spontaneous speech recognition.

Acknowledgements

The work was supported with Russian Fund of Basic Research, project 14-01-00607

Bibliography

[Jelinek, 1997] Jelinek F. Statistical Methods for Speech Recognition. Cambridge, Massachusetts: The MIT Press, 1997.

[Wester, 2003] Wester M. Pronunciation modeling for ASR âĂŞ knowledge-based and data-derived methods // Computer Speech and Language. 2003. Vol. 17, P. 69-85.

[Fosler, 1999] Fosler-Lussier E. Dynamic pronunciation models for automatic speech recognition. Ph.D. thesis. University of California, Berkley, CA, 1999.

[Saraclar, 2004] Saraclar M., Khudanpur S. Pronunciation change in conversational speech and its implications for automatic speech recognition // Computer Speech and Language. 2004. Vol. 18(4). P. 375-395.

[Saraclar at al., 2000] Saraclar M., Nock H., Khudanpur S. Pronunciation modeling by sharing Gaussian densities across phonetic models // Computer Speech and Language. 2000. Vol. 14(4). P. 137-160.

[Zheng, 2003] Zheng J., Franco H., Stolcke A. Modeling word-level rate-of-speech variation in large vocabulary conversational speech recognition // Speech Communication. 2003. Vol. 41. P. 273âĂŞ285.

[Bahl, 1983] Bahl . R., Jelinek F., Mercer R. L. A maximum likilihood approach to continuous speech recognition // IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol 5. 1983. P. 179-190.

[Schwartz, 1990] Chow Y.-L., Schwartz R. The N-Best Algorithm: Efficient Procedure for Finding Top N Sentence Hypotheses // Proceedings of the International Conference on Acoustic, Speech and Signal Processing, ICASSP. 1990, P. 199-202.

[Young, 1997] Young S., Bloothooft G. (Eds.). Corpus-based methods in language and speech processing. Dordrecht: Kluwer Academic Publishers (Text, Speech and Language Technology series, Vol. 2), 1997.

[Bogdanov, 2004] Bogdanov D. S., Krivnova O. F., Podrabinovitch A. J., Arlazarov V;L. Creation of Russian Speech Databases: Design, Processing, Development Tools // Proceedings of the International Conference on Speech and Computers, SPECOM. Moscow, 2004.

[Chuchupal, 2005] Chuchupal V.J., Makovkin K.A., Chichagov A.V., Kuszetsov V.B., Ogarysyev V.F. Speech corpora TeCoRus. Data base registration sertificate 2005620205, 2005.

[Echo Moscow] http://www.echo.msk.ru.

Authors' Information

Vladimir Chuchupal *- Leading scientist, Dorodnicyn Computing Centre Russian Academy of Sciences, P.O. Box: 111333, Moscow, Vavilov Str, 40, Russia; e-mail: v.chuchupal@gmail.com*
Major Fields of Scientific Research: Speech recognition, signal processing,

Anton Korenchikov *- Student, Lomonosov Moscow State University, P.O. Box: 111333, Moscow, Vavilov Str, 40, Russia; e-mail:*
Major Fields of Scientific Research: Speech recognition

TABLE OF CONTENTS OF IJ ICP VOL.1, NO.:1, 2014

TABLE OF CONTENTS OF IJ ICP VOL.1, NO.:2, 2014

TABLE OF CONTENTS OF IJ ICP VOL.1, NO.:3, 2014

TABLE OF CONTENTS OF IJ ICP VOL.1, NO.:4, 2014