SEMANTIC ENCODING AND MARKUP OF GEOREFERENCED DOCUMENTS IN POLYTHEMATIC DIGITAL LIBRARIES OF SCIENTIFIC LITERATURE [1)]

Igor M. Zatsman
Institute for Informatics Problems of the Russian Academy of Sciences,
2 build., 44, Vavilova st., Moscow, 117333, Russia,
E-mail: zatsman@bur.oivta.ru

Abstract. The paper considers the principles and basic stages of decomposing georeferenced documents oriented to the problems of markup and semantic search. The paper justifies the necessity to develop a multimodal semiotic system and discusses verbal-visual knowledge representation in digital libraries. To represent knowledge, verbal-visual thesaurus is proposed. The thesaurus includes verbal, verbal-visual and visual descriptors. Semiotic synonymy is included into the system of links between descriptors in the thesaurus. The concept of generalized semantic search of documents in an digital library is defined. It is proved that semiotic synonymy in the thesaurus provides for cognitive framework for solving the problems of generalized semantic search of documents.

1 Introduction

To develop new technologies for creation and use of digital libraries that incorporate text, images, audio, and video, efforts are taken to implement semantic search and retrieval. New technologies will be based on the following classes of methods and algorithms, which have been examined and subjected to experimentation in various digital libraries [i]:

- object recognition, segmentation, and indexing,
- semantic analysis,
- knowledge representations,
- human-computer interactions and information visualization.

In this paper, polythematic digital libraries are inspected for documents in Earth sciences (hereinafter referred to as "georeferenced documents"). Beside the plain text, these documents may contain maps, schemes, plans, charts, airphotos, spacephotos, map-type diagrams, etc. In georeferenced documents, non-verbal components can be even more valuable than their verbal components. It accounts for the facts that scientific information contained in non-verbal components may be absent in the plain text of a scientific document. At the moment, to integrate full-text georeferenced documents in polythematic digital libraries maps and geospatial footprints are encoded, as a rule, as simple, i.e., non-structured bit patterns. Sometimes, figure captions are isolated to be included into search area together with plain text of a document [ii].

From the above-listed classes of methods and algorithms, only visualization has been implemented for maps and geospatial footprints in digital libraries of scientific literature. Recognition, decomposition, segmentation, markup, indexing and semantic search of map and geospatial footprint objects are at the stage of formulation [iii, iv].

The paper focuses on decomposition, semantic encoding, and markup of georeferenced documents. To consider the items, let us specify the terms and concepts used in the paper.

Verbal document components - linear discrete concatenations of characters. These are natural language fragments of the title, abstract, sections, chapters, paragraphs, figure captions as well as the natural language text in charts, maps, schemes, graphs, tables, etc.

Structural components - multilinear discrete concatenations of characters or/and signs as well as their combinations joined together by network, hierarchical, relational, and discrete parametrical schemes. These are mathematical formulae, structural chemical formulae and reactions, bioinformatic sequences, etc.

Graphical components - continuous or discrete-continuous combinations of graphical signs which can be static or dynamic, one-dimensional or multi-dimensional. This component category also includes signs with fuzzy, random or uncertain boundaries and forms. These are graphs, charts, schemes, sketches, maps, drawings and photos as well as animation and dynamic information objects of georeferenced documents.

[1)] The work is supported by the Russian Foundation for Basic Research, project 99-05-65491.

Document components can be embedded. For example, a table can have text cells, cells with chemical formulae as well as cells with images. Embedding and/or various combinations of the three above-mentioned types of *homogeneous components*, i.e. verbal, structural, graphical, provide for four types of *heterogeneous components*, i.e. verbal-structural, verbal-graphical, structural-graphical and verbal-structural-graphical. To designate structural, graphical as well as any types of heterogeneous components, the term "non-verbal" components can be used.

Traditional (paper, conventional) documents include verbal and non-verbal components with certain, static and one-dimensional or two-dimensional signs only. *Generalized documents* include verbal and non-verbal components with fuzzy, random or uncertain or/and dynamic or/and three(multi)-dimensional signs.

Semantic search can be defined as search based on semantics of all the document components with the three basic types of meaning expression, i.e. presentational, organizational and orientational [v].

Sign modality means that the sign belongs to some language or language system: verbal modality denotes the category of natural languages; mathematical modality stands for mathematical formulae, chemical modality stands for chemical formulae, etc. [5]. Signs that belong to hetero-modal languages are called *polymodal*. Semiotic systems with signs of different modalities, polymodal signs included, are referred to as *multimodal*. The core of the proposed cognitive framework for digital libraries is the *multimodal semiotic system* integrating verbal and non-verbal signs.

Sometimes, the data that the user of the digital library is interested in can be represented in different modalities. In that case the user can not know the presentation form of the target data in the digital library (verbal, structural, graphical, verbal-structural, verbal-graphical, structural-graphical or verbal-structural-graphical). Therefore, the query modality may be different from that of data representation in digital libraries. In case the modality is not known in advance and the query contains any of the possible modalities are referred to as *generalized semantic search*.

If the authors of the document define contents of signs which are used only by them these signs are referred to as *author's (proprietary) signs*. Signs which forms and contents do not change for a long time are referred to as *definite*. In case the used signs are not universally adopted but are used by many authors are referred to as *semidefinite*. When speaking of scientific documents both traditional (paper, conventional) and generalized documents are meant.

2 Decomposition of full-text georeferenced documents

Introduction to semiotic and semantic problems of modelling and search of full-text scientific documents in digital libraries can be found in [vi, vii]. These papers also feature the description of the first two stages of logical and semantic modeling of documents.

At the first stage the basic problem is to decompose a scientific document into homogeneous and heterogeneous components. The first stage should result in a structural model of the document. All the employed schemes of structurization should be indicated. All the addressable homogeneous and heterogeneous document components are listed in the model. At present methods for the initial stage of modeling are developed within the framework of numerous application projects [viii].

At the first stage of logical and semantic modeling content-based characteristics of georeferenced documents and their components can be taken into account. The stage of decomposing georeferenced documents and their components would be oriented to organization of semantic search in digital libraries. The incapability of conventional models (hierarchical, network, and relational) to model, store and process the whole content of a document in an appropriate way has been recognized in respect to semantic search and retrieval [ix, x].

Four schemes to describe relations and links that reflect a multilevel embedded structure if scientific documents and their components were considered, i.e., hierarchical, network, relational, and parametric schemes. It was shown that sometimes it is necessary to perform parameterization based on multiple continuous and discrete arguments. To perform parameterization, well-known functions of multiple continuous and discrete arguments are proposed. The spatiotemporal scheme is considered as a special case of the parametric scheme [7, xi].

For georeferenced documents parametric and spatiotemporal schemes provide for a more precise decomposition, segmentation and content representation. That can be applied both to electronic forms of conventional documents and generalized documents.

Let us formulate the principles of full-text georeferenced document decomposition in digital libraries bearing in mind the examples and research results [7, 10]:

- multilevel embedding and/or various combinations of verbal, structural and graphical components;
- a combination of several schemes to describe links and relations within one document (for a general case, a combination of hierarchical, network, relational, and parametric);
- representation of a single document in multiple electronic forms, based on different backbone schemes [11];
- the conventional modern frame of longitude reference; shift of paleocoordinates depends upon the geological time.

The first three principles can be referred to any scientific documents while the fourth principle deals with content-based matters in Earth sciences. In particular, when paleotectonic footprints are parameterized the shift of the paleoequator and poles should be taken into account.

The second stage of semantic document modeling includes the semantic mark-up and encoding of document components as sign representations. Paper [10] shows that current semantic mark-up languages are mainly oriented to verbal components and some types of structural components. These languages cannot be used for mark-up of a wide spectrum of graphical data. To provide for semantic encoding of graphical components, the mark-up languages should be upgraded. One of the possible approaches to upgrading could be the introduction of constructions for sign representations of continuous images. To obtain sign representations, the semiotic approximation method can be applied.

3 Semiotic approximation method

In [6] the generalized concept of sign is proposed with the understanding that sign representations of digital forms of scientific documents should be obtained. The proposed generalized concept of sign provides for a unified semiotic basis for obtaining sign representations both for verbal and graphical components of documents. Sign representations of graphical components are characterized by the following features:

- multiple variants of sign representation of graphical components,
- indefinite language belonging of individual signs,
- indefinite contents of proprietary and semidefinite signs.

The first feature testifies to the fact that for a single graphical component multiple sign representations can be obtained. Graphical signs should be normalized. The procedure of normalizing signs of graphical components differs from that of verbal components. The essence is the considerably larger number of non-normalized variants of graphical signs as compared with verbal ones. In theory, the number of variants is infinite [6].

The central idea of the paper is the use of the verbal-visual thesaurus to solve the following problems:

- normalization of signs for semantic encoding of graphical components of full-text georeferenced documents;
- explicit definition of language belonging of signs;
- contextual definition of proprietary and semidefinite signs.

Sign representations of graphical components can be obtained with the help of combinations of normalized graphical signs referred to as descriptors of the verbal-visual thesaurus. We suggest that this solution should be called the semiotic approximation method.

In accordance with the suggested typology of document components a multimodal semiotic system of a polythematic digital library must include the following basic parts:

- traditional verbal sign systems of natural languages,

- structural sign systems,
- graphical sign systems,
- heterogeneous sign systems (verbal-structural, verbal-graphical, structural-graphical and verbal-structural-graphical).

The present paper deals with the principles of design of graphical sign systems. Graphical signs which are necessary for semantic encoding of graphical components of full-text georeferenced documents can be chosen with the help of semiotic approximation method.

To describe the method, let us define the following concepts: sign-set, metasign, semiotic approximation, sign basis, verbal-visual thesaurus. Their semantics is the essence of the proposed method.

The concept of sign-set is introduced to differentiate between the generalized sign and the traditional one which is adopted in "A Theory of Semiotics" [xii]. Within the framework of the generalized interpretation the forms of signs and their combinations can be [6]:

- certain or uncertain (fuzzy, random, multiversion),
- static or dynamic,
- one-dimensional or multi-dimensional,
- discrete, continuous or discrete-continuous.

Thus the concept of sign-set will be used to cover the above-mentioned spectrum of sign form features. In terms of mathematics, sign-sets can be analyzed with the help of the set theory while formal logic procedures can also be applied. In terms of semiotics, sign-set can be referred to as non-verbal signs and their contents can be used for knowledge representation in digital libraries.

Sign-sets with multi-dimensional forms are very difficult to include into traditional articles. For example, in digital climatic maps, dynamic, three-dimensional and fuzzy sign-sets might be convenient to represent continuous change in cloudiness for some period. However, in traditional editions, this is next to impossible.

To designate sign-sets in traditional documents, the corresponding metasigns can be used in the form of indexed letters of the alphabet. In computer information processing, metasigns can be also used as indicators of the place where sign-sets are stored. The relationship between the metasign, sign-set, referent, and concept can be schematically reproduced as a tetrahedron and is referred to as a generalization of the semiotic triangle by Frege.

Semiotic approximation is the sampling of graphical components with the use of normalized sign-sets and their combinations which are constructed in advance. The combination of normalized sign-sets produced for any category of homogeneous (in terms of theme and object) graphical components of full-text georeferenced documents on the basis of the verbal-visual thesaurus and its descriptors can be referred to as the *sign basis* of this category. It should be stressed that sign-sets reflect the semantics and spatial features of objects of all the graphical components of this category that are stored in the digital library.

Descriptors of the verbal-visual thesaurus are used as normalized signs for semantic encoding of graphical components. Thus, in the digital library, the thesaurus is an authority of sign basis of all the categories of graphical document components. As a first approximation, assume that the verbal-visual thesaurus is a thesaurus which integrates semantic description of verbal, structural and graphical information of the digital library.

Currently, the semiotic approximation method of graphical components is at the initial stage. Thus the application scheme of the method is preliminary. First, for each category of graphical components of full-text georeferenced documents in the digital library sign basis are constructed. Second, at the stage of semantic encoding of some graphical component the thematic and object genus of the component is defined. In case of thematic and object homogeneity of the component sign-sets are extracted from the corresponding sign basis. The sign-set is extracted if their semantics and spatial characteristics correspond to the objects of the graphical component.

The combination of extracted sign-sets can be considered as normalized sign representation of the graphical component. In case of thematic and object heterogeneity of graphical components the combinations of sign-sets from different basis can be used. As it follows from the definition of sign basis, their authority is the verbal-visual thesaurus.

4 Typology of knowledge representation and verbal-visual thesaurus

This thesaurus is constructed on the basis of the following typology of knowledge representation. There are three basic spheres of knowledge representation [12]:

- verbal knowledge in linguistic form which cannot be adequately converted in a non-verbal form (Sphere I),
- non-verbal (non-linguistic) knowledge which cannot be represented in the verbal form (Sphere II),
- knowledge that is adequately represented neither in verbal nor non-verbal form.

The relationship between the three spheres of knowledge representation can be metaphorically illustrated as two partially intersected circles (Fig. 1a). The first circle designates the knowledge representation sphere in the linguistic form, while the second circle stands for the knowledge representation sphere in the non-linguistic form. The area at the intersection of these circles corresponds to the third knowledge representation sphere.

Let us point out that in the set theory the example of two intersected circles illustrates the intersection of two sets, i.e., the elements that belong to two sets. However, in Fig. 1a, the intersection of two circles denotes the elements of the third set which do not belong to the first two sets.

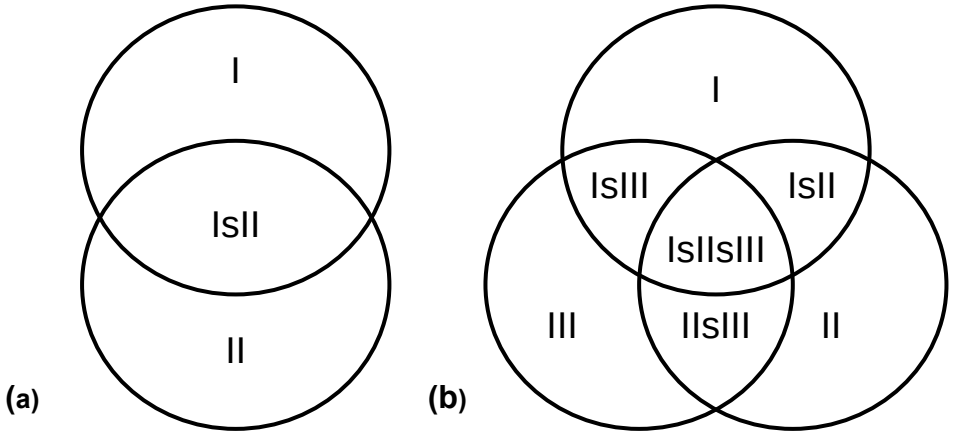

Fig. 1. Knowledge representation spheres in linguistic and non-linguistic forms (a); in verbal, structural and graphical forms [seven from eleven possible spheres are designated] (b).

Therefore, Fig. 1 should be considered in terms of semiotics in accordance with the typology of knowledge representation spheres listed in [12] and not from the point of view of the set theory. To designate the third sphere of knowledge representation which can be referred to as the area of semiotic synonymy, let us use I, II, and the letter "s" which stands for "synonymy".

There are a great variety of non-linguistic forms of knowledge representation. The present paper discusses structural and graphical components of scientific documents as well as their combinations with verbal components. Thus, a more detailed typology of knowledge represented in digital libraries is proposed. Let us consider seven knowledge representation spheres which are charted in Fig. 1b as seven areas formed by three intersected circles. The three circles correspond to verbal, structural, and graphical forms of knowledge representation. Let us point out that Fig. 1a features only one sphere of semiotic synonymy. Four spheres of semiotic synonymy labelled as IsII, IsIII, IIsIII, IsIIsIII in Fig. 1b are obtained in case of a detailed typology and classification of knowledge representation spheres as verbal (I), structural (II), and graphical (III).

For example, segment IsIIsIII corresponds to the part of semiotic synonymy where knowledge can be adequately represented in any of the three forms: verbal, structural and graphical. In contrast to the three possible variants of representing

the same knowledge in segment IsIIsIII semantics of verbal-structural-graphical components is represented by the simultaneous combination of the verbal, the structural and the graphical. This part of knowledge is not shown in Fig. 1b.

Thus Fig. 1b features seven spheres of knowledge representation but other four spheres are missing. They correspond to four types of heterogeneous components of documents in digital library. To demonstrate these four spheres of knowledge representation, let us depart from the visual metaphor of knowledge typology described in [12]: the position of IsII in the center of Fig. 1a. Let us diagram all the spheres of semiotic synonymy outside the circles I, II, III (Fig. 2).

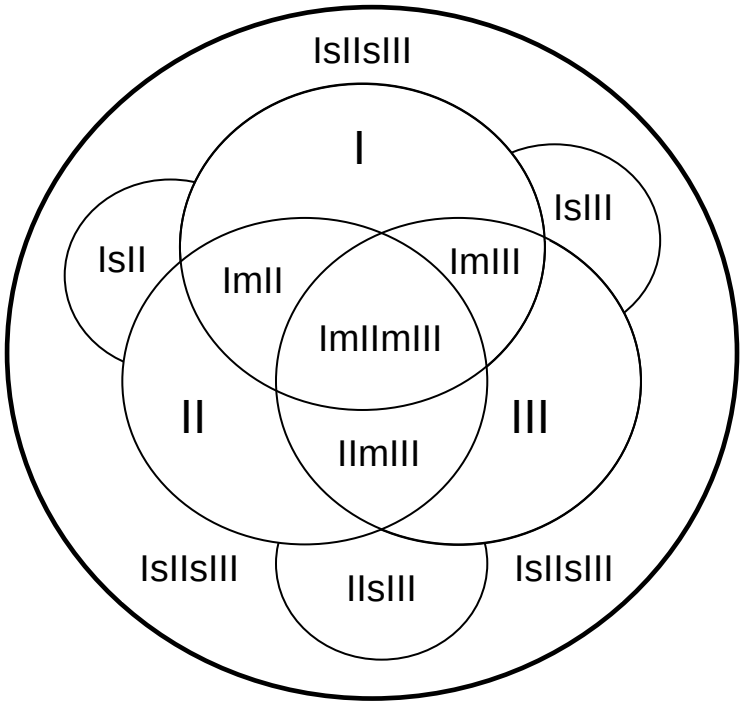

Fig. 2. Knowledge representation spheres in digital libraries of full-text scientific documents

Fig. 2 features the eleven spheres of knowledge representation. Thus four areas ImII, ImIII, IImIII and ImIImIII can be sketched. These areas which, by analogy with heterogeneous components, can be referred to as verbal-structural, verbal-graphical, structural-graphical, and verbal-structural-graphical knowledge representation spheres. The names of the spheres contain the letter "m" which stands for miscellaneous. The typology of eleven knowledge representation spheres reflect the cognitive structure of the semantic space of the digital library of full-text scientific documents with four areas of semiotic synonymy included.

The proposed typology of knowledge representation spheres will be used as the basis for describing the concept of verbal-visual thesaurus. The seven spheres of knowledge representation labelled as I, II, III, ImII, ImIII, IImIII, and ImIImIII can be correlated to seven types of homogeneous and heterogeneous descriptors of the verbal-visual thesaurus.

To specify verbal-structural, verbal-graphical, and verbal-structural-graphical descriptors, let us use the concept of verbal-visual descriptors. To designate structural, graphical, and structural-graphical descriptors, we shall use the concept of visual descriptors.

In accordance with these definitions the verbal-visual thesaurus includes four spheres of semiotic synonymy labelled as IsII, IsIII, IIsIII, IsIIsIII and seven spheres labelled as I, II, III, ImII, ImIII, IImIII and ImIImIII, with each sphere containing the corresponding category of descriptors.

For example, in the graphical sphere labelled as III graphical descriptors are placed. This section will provide for graphical sign-sets to build sign basis for specific classes of graphical components.

The inclusion of visual and verbal-visual descriptors into the thesaurus is accompanied by the extension of link types between the descriptors. An entirely new type of link "semiotic synonymy" appears. Figures 1b and 2 feature four knowledge representation spheres with this type of link.

Only these sections of the thesaurus make use of this type of link between the descriptors which differ in knowledge representation forms and language modalities but stand for the same concept. This is the difference between semiotic synonymous descriptors from traditional verbal synonymous descriptors. The latter refers to one verbal knowledge representation sphere and have identical verbal modality.

Let us consider semiotic synonymy illustrated by the example of international chemistry nomenclature which is a set of names of chemical substances, their groups and classes as well as the rules of nomination. To make up a name based on a structural chemical formulae or to convert a name into a chemical formula, first a set of formal rules is performed. In accordance with nomination rules for organic substances a type structure with the adjacent basic compound group is revealed and nominated. To make names, the classification of compounds and the names of characteristic groups of organic compounds are used.

In this example, the role of synonymous descriptors is played by the structural chemical formula and the name of the chemical compound if they are included into the thesaurus as a structural descriptor and a verbal descriptor. Both structural and verbal synonymous descriptors logically belong to the sphere of the thesaurus labelled as IsII. Let us point out that in a scientific document the corresponding concept can be designated by verbal or structural signs.

The link of semiotic synonymy in the verbal-visual thesaurus provides for the generalized semantic search in digital libraries. The data that the user of the digital library is interested in can be represented in different modalities. If the digital library has the verbal-visual thesaurus with the link of semiotic synonymy the query can contain any of the possible modalities. Therefore, the query modality may be different from that of data representation in digital libraries.

5 Discussion

The semiotic approximation method is proposed as a basis for solving the problem of normalization of sign representation of graphical components. This method allows one to choose normalized graphical signs for semantic encoding of graphical components of documents. Normalized signs are selected from the verbal-visual thesaurus of the digital library.

Aside from normalization, the thesaurus can be used for contextual definition of proprietary and semidefinite signs of document components. For contextual definition of proprietary signs in some document a correspondence between signs forms and descriptor contents is established. To identify the corresponding descriptors of the thesaurus, the content of document components is used.

A different procedure is proposed for semidefinite signs. While proprietary sign contents are described in the document the semantics of semidefinite signs is not often disclosed. Thus [6] suggests that peculiarities of semidefinite signs should be taken into consideration when the digital library is created. The corpora of scientific documents should be supplied with materials describing the used semidefinite signs.

The necessary descriptors to build up sign basis and/or contextual definition of proprietary and semidefinite signs might be missing in the thesaurus. Such a situation is also possible when traditional verbal thesauri are supported. In this situation the thesaurus must be expanded. When documents with proprietary signs are processed the contents of the documents can be used to expand the verbal-visual thesaurus with needed descriptors. When documents with semidefinite signs are processed the materials describing their semantics can be made use of.

The sign basis created from descriptors inherits the unity of the form and the content expressed in the system of thesaurus links as well as their language belonging. When a document component includes individual signs with indefinite language property the language belonging of the signs can be fixed with the help of the corresponding descriptors. Since for each descriptor its language belonging is explicitly expressed signs of the basis for this component have definite language property.

6 Conclusions

Aside from generalized semantic search in digital libraries the multimodal semiotic system is of special importance for further development of the XML metalanguage. One of the possible directions of its development is associated with the multimodal semiotic system and the verbal-visual thesaurus.

When documents with graphical components are marked-up we propose to use the new mark-up language constructions that include metasigns as references to specific descriptors in the thesaurus. When documents are processed by the computer the metasigns are used as indicators to storage places of descriptors in the thesaurus. Due to the introduction of metasigns and new constructions of the semantic mark-up languages their application can be extended to graphical components of full-text scientific documents.

The sampling of graphical components with the use of sign bases is approximate. The descriptors of the verbal-visual thesaurus which are used for the semantic encoding of graphical component are not identical to fragments of this component. Therefore, in semantic mark-up languages, metasigns can be used as indicators to sign-sets in order to solve the problems of search and retrieval. To visualize and reproduce graphical components of searched documents, raster or vector digital forms of graphical components can be used.

References

1. Chen H. Semantic Research for Digital Libraries // D-Lib Magazine. -1999. -Vol. 5, N 10.

2. Schatz B., Mischo W., Cole T., Bishop A. et al. Federated Search of Scientific Literature // Computer. - 1999. -Vol. 32, N 2 - P. 51-59.

iii. Goodchild M.F. Implementing Digital Earth: A research Agenda // Proceedings of the International Symposium on Digital Earth. - Beijing: Science Press, 1999. - pp. 21-32.

iv. Di L., McDonald K. Next Generation Data and Information Systems for Earth Sciences Research // Proceedings of the International Symposium on Digital Earth. - Beijing: Science Press, 1999. - pp. 92-101.

5. Lemke J.L. Multiplying Meaning: Visual and Verbal Semiotics in Scientific Text. In: Martin J.R. and Veel R. (Eds.) Reading science: Critical and functional perspectives on discourse of science. - London: Routledge, 1998. - pp. 87-113.

6. Zatsman, I.M. Semiotic Problems of Modelling and Search of Full-text Scientific Documents // Trudy mezhdunarodnogo seminara Dialog-2001 po kompyuternoi lingvistike i ee prilozheniyam (Proceedings. Dialogue-2001 International Workshop "Computational Linguistics and its Applications"). - Aksakovo, 2001. - pp. 136-144 (in Russian).

7. Zatsman I.M. Logiko-semanticheskiye modeli polnotekstovykh nauchnykh dokumentov (Logical Semantic Models of Full-text Scientific Documents) // Nauchno-tekhnicheskaya informatsiya (seriya 2), 1999, No. 5, pp. 13-22 (in Russian).

8. Lee K.H., Choy Y.C., Cho S.B. Geometric Structure Analysis of Document Images: A Knowledge-Based Approach // IEEE Transactions on Pattern Analysis and Machine Intelligence. - 2000. - Vol. 22, No. 11, pp. 1224-1239.

ix. Jarvelin K., Niemi T. Integration of complex objects and transitive relationships for information retrieval // Information Processing & Management. - 1999. - Vol. 35 - P. 655-678.

x. Zatsman I.M. Strukturizatsiya i semanticheskaya razmetka dokumentov v elektronnykh bibliotekakh (Structuring and Semantic Markup of Documents in Digital Libraries) // Trudy mezhdunarodnogo seminara Dialog-1999 po kompyuternoi lingvistike i ee prilozheniyam (Proceedings. Dialogue-1999 International Workshop "Computational Linguistics and its Applications"), vol. 2. - Tarusa, 1999. - pp. 67-73 (in Russian).

xi. Zatsman I.M. Electronnye kollektsii polnotekstovyh nauchnyh documentov (Electronic Collections of Full-text Scientific Documents) // Sistemy i sredstva informatiki. Issue 9. - Moscow: Nauka, 1999. - pp. 177-202 (in Russian).

12. Eco U. A Theory of Semiotics. - Bloomington: Indiana University Press, 1976. - 356 pp.