

USING THE THESAURUS TO DEVELOP IT INQUIRY SYSTEMS

FEDOTOV A.M.¹, TUSUPOV J.A.², SAMBETBAYEVA M.A.², SAGNAYEVA S.K.²,
BAPANOV A.A.², NURGULZHANOVA A.N.³, YERIMBETOVA A.S.²

¹Institute of Computational Technologies, Siberian Branch of the RAS, Novosibirsk, Russian Federation

²L.N. Gumilyov Eurasian National University, Astana, Kazakhstan

³M.Tynyshpaev Kazakh Academy of Transport and Communications, Almaty, Kazakhstan

E-mail: aigerian@mail.ru, sagnaeva_tar@mail.ru, madina_jgtu@mail.ru, fedotov@ict.nsc.ru,
tussupov@mail.ru

ABSTRACT

The article describes the standards of thesauruses, as well as their possible usage in scientific and educational information systems. The authors compare the various approaches to the description of the data schema based on the object model. Particular attention is paid to SKOS and Zthes data schemes, and the second one has been selected for implementation. The study also considered the work with dictionaries of key terms used for the organization and classification of information resources.

Keywords: *Z39.50, data schema, SKOS, Zthes, MARC, VDEX, MODS, information system, electronic library, dictionary, directory, database, information retrieval thesaurus, metadata.*

1. INTRODUCTION

The development of information technology in general, as well as in the field of communication and information processing, in particular, has led to the emergence of fundamentally new opportunities for organizing virtually all the phases of the scientific and educational process that, in its turn, has led to the qualitative growth of the informational needs of scientists and teachers. A modern student armed with a computer, using everyday possibilities of the Internet, cannot be satisfied with the traditional mode of teaching process and conventional formats of educational materials such as textbooks, books, or flat text files. Today, training materials can be provided to students in a variety of digital formats, and they must be supported by a variety of search and classification services. Systematization and classification of existing information resources in accordance with the user's needs is one of the most important tasks to support both scientific and educational activities [1, 2].

In the process of scientific and especially educational activities, much time and energy is taken by the work with literary sources, various types of materials and documents: the search for

the necessary documents, the systematization and classification of documents in accordance with the task. In order to satisfy information needs of today's users, support for complex retrieval functions and classification of the information are required, as well as viewing of the resources by categories (headings) and by dictionary-classifiers. The most important task is to systematize the resources, for the solution of which it is necessary to clearly define the structure of logical-semantic categories (facets) and the key terms (concepts) that cover the chosen narrow subject area that is interesting to the user. Usually the subject area is limited by training courses or specific topic of the course.

At present time, there are enough powerful information systems, which to a greater or lesser degree satisfy the information needs of the users [3]. However, the main disadvantages of the number of systems is the limited possibility of carrying out the analytical work with the resources and providing the integration of the resources both inside of a system and in external systems (the international standards and recommendations are often not taken into account, low interoperability) [4]. This is extremely inconvenient in the field of scientific and educational activities, one of the



main tasks consists in the fact, that it is necessary to establish links between specific scientific facts (for example, "What is meant by cybernetics" or "who is the author of this article") and entities of an information system (person, facts, documents, publications, etc.).

The standard approach to systematization of information is the classification of documents by using taxonomies. Taxonomy is a subject classification, which groups terms by the way of controlled vocabulary (thesaurus) and organizes them (dictionaries) as a hierarchical structure. To describe any subject area a certain set of key terms is generally used, each of which represents or describes any notion of this subject area [5]. The basis of classification makes the selection of concepts (key terms), the establishment of paradigmatic relations (for example, parent - child type) between them and comparison of the analyzed document by the selected concepts.

The most inconvenient thing in providing the educational information systems lies in the fact that the technologies of classification and systematization of information developed by libraries during the last hundred years do not work [6], as a result of the thematic proximity of the classified documents. For example, the most suitable for classification of resources in mathematics or informatics UDC¹ dictionaries and MSC2000² or UNESCO³ thesaurus, usually classify all the resources selected for a particular training course to one category.

In monograph [7], published by VINITI in 1976 and containing a detailed overview of the theoretical problems of the information retrieval, based on the selection of two types of information needs - the need for information about the sources of necessary scientific information and the need for necessary scientific information itself - states, that to meet information needs of the first one the information systems are designed, which have received the name – the documentary, for the second type – factographic. At present, the most required tools of the information support of

scientific activities are informational systems combining the capabilities of information systems of the both above mentioned types and which allows satisfying the information needs of a qualified user in accordance with the scheme "document - fact - classification" [4, 8]. In such informational systems, the deliberative system is as an integral component, formalizing the rules of logical inference, and work with facts and concepts characterizing a particular document.

2. THESAURUSES IN INFORMATION RETRIEVAL

According to the definition of the International Organization for Standardization (ISO), a thesaurus is a dictionary, controlled by indexing language, formally organized in order to establish a clear priori relations between the concepts [9, 10]. This definition sets lexical units and semantic relations between these units as the elements that make up the thesaurus. Thesaurus relations (genus - species, part -whole, complex - element, cause - consequence) are imposed on the structure of the taxonomy, i.e. basic taxonomies of the subject area are identified.

Historically thesauruses were designed for manual indexing of documents and when they were being created the issues related to the automatic indexation were not taken into account. The difficulty of making a thesaurus, corresponding to all the thematic variety of information, which meant to be indexed, is the primary cause of its unpopularity in the modern information systems. However, as we have already noted, the effectiveness of information retrieval systems for support scientific and educational activities directly depends on the use of specialized thesaurus.

One of the first thesauruses in the history and the most famous for today is the thesaurus compiled by the British lexicographer Peter Mark Roget, published in 1852. The original name of the Roget's Thesaurus is "Thesaurus of English Words and Phrases".

In connection with computing machines, for the first time a thesaurus was used by M. Masterman in 1954 in the field of machine translation [11]. By using the thesaurus the correspondence between language of user's requests and documents in the information system was set. Ju. A Schrader [12] proposed to consider a thesaurus as a system of knowledge, reflected by the language, and then

¹ Universal Decimal Classification, supported by the International Federation of Documentation (Federation Internationale de Documentation - FID) and the UDC Consortium <http://www.udcc.org/>. The Russian version is supported by the All-Union Institute of Scientific and Technical Information of the Russian Academy of Sciences.

² Mathematics Subject Classification (<http://www.ams.org/msc/>), supported by the American Mathematic Society (AMS).

³ <http://databases.unesco.org/thesru/>



thesaurus became interesting in itself, not only as an auxiliary tool.

Among the universal automated thesauruses there should be noted an intelligent computer thesaurus WORDNET⁴ (vocabulary of the English language), a similar one is RussNet⁵ for the Russian language, and the aforementioned multilingual UNESCO thesaurus.

As the specialized thesauruses, which have already been implemented media-resident with software support for users we can designate: EuroVOC⁶ - large retrieval thesaurus used for indexing documents of the European Union; thesauruses and controlled vocabularies of the Research Service of the US Library of Congress⁷; thesauruses of the American Society for Indexing ASI⁸; known thesaurus AGROVOC⁹, covering the terminology of Agriculture and Forestry; computerized thesaurus of the medical terminology SNOMED¹⁰, and others.

Among the Russian specialized thesauruses there should be noted the thesaurus of the Central Scientific Agricultural Library in the field of agriculture¹¹. A thesaurus constructed as an extension of the SRSTI¹² dictionary-classifier and format description terms compatible with AGROVOC.

3. THESAURUSES IN THE DESCRIPTION OF INFORMATION

Information retrieval thesaurus (in accordance with the definitions of the standards) is a normative (controlled) dictionary of key terms in a natural language with explicit semantic relations between terms, supposed to describe the content of the documents and search queries [13, 14]. Thesaurus is intended to describe a particular subject area, each term which denotes or describes any notion of this subject area.

Terms, which are the constituents of the thesaurus, are classified into descriptors (authorized terms) and non-descriptors (ascriptors). Descriptors definitively correspond to the concepts of the subject area. The relations between descriptors are usually divided into two types: hierarchical and associative. Hierarchical relations are generally regarded as asymmetric and transitive.

In accordance with GOST 7.25 - 2001 [15] the hierarchical relations have properties of transitivity and antisymmetry, which can be used in excessive indexing in order to improve efficiency of information search.

Applied in the information retrieval thesauruses, hierarchical relations can be differentiated into separate types. The primary hierarchical relations used in information retrieval thesauruses, is the genus-species relation (parent-child, wider-narrower, above-below). Genus-species relation is established between the two descriptors, if the volume of concept of the lower descriptor is included in scope of the concept of a higher descriptor. Also, as the hierarchical relation in information retrieval thesauruses we can establish part-whole relations. Relation of association is non-hierarchical and associative. The main purpose of establishing the associative relations between descriptors of the information search thesaurus is the indication of the additional descriptors, which are useful in indexing or search [17].

It should be noted that the model of the information retrieval thesaurus described in the national and international standards is intended for using it in the process of manual, expert analysis of documents [13 - 16].

The main purpose of the development of the traditional information retrieval thesauruses is to use their units (descriptors) to describe basic topics of documents in the process of manual indexing. So it is important, that a set of descriptors of the information search thesaurus should allow describing subjects of the documents of the subject area. At the same time, the indexing process itself in such thesaurus rests on linguistic, grammatical knowledge, as well as the knowledge of the subject area which the professional text indexers have. The indexer first should read the text, understand it and then to state the content of the text, using the descriptors, specified in the informational retrieval thesaurus. The indexer must understand all the terminology used in the text, thus for the

⁴ <https://wordnet.princeton.edu/>

⁵ <http://project.phil.pu.ru/RussNet/>

⁶ <http://europa.eu/eurovoc/>

⁷ <http://www.loc.gov/>

⁸ ASI - American Society for Indexing
<http://www.asindexing.org/>

⁹ <http://www.fao.org/agrovoc/>

¹⁰ <http://www.ihtsdo.org/>

¹¹ <http://www.cnshb.ru/>

¹² State Rubricator of Scientific and Technical Information
<http://gnti.ru/>

description of the main theme of the text he will need much fewer terms [5, 17].

4. STANDARDS OF THESAURUS PRESENTATION

A series of international (ISO), national and corporate standards and recommendations are developed for presentation of thesaurus. Figure 1 shows their evolution.

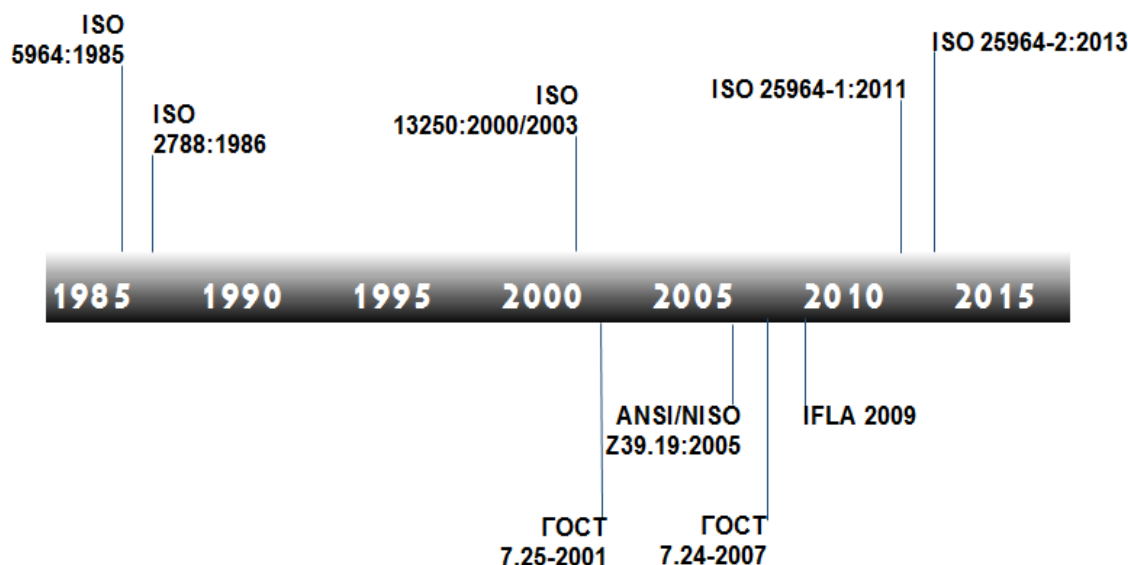


Figure 1. The Evolution of Standards and Recommendations for Presentation of Thesaurus

The standard entitled "Guidelines for the Establishment and Development of a Multilingual Thesaurus", ISO 5964: 1985 [9] defines the use of different languages with application of the data, pointing to the main questions regarding the semantic equivalence.

The ISO 2788: 1986 standard [10] for monolingual thesauruses ("Guidance for Establishment and Development of Monolingual Thesaurus"), first published in 1974 was revised and published in 1986. This standard of the international level had been actual for 25 years before the introduction of standard ISO 25964-1 [13] in 2011.

The ISO 13250 standard [18 - 20] defines the thematic maps (TM). The information is presented in xml format with explanations to xmt designations. In the standard of xmt format, key elements and concepts are shown (topics, associations, origin, published themes and field of application). Also TM standard typically includes format of electronic interchange thesauruses, although it does not separately identifies for them, because, it is one of its applications.

The American National Standards Institute (ANSI) has revised the standards relating to the thesaurus under the influence of the Guidance on Creation, Format and Management of Monolingual Controlled Vocabularies. It offered guidance for creating a thesaurus with an emphasis on the format and control of monolingual dictionaries. Recommendations of the ANSI / NISO Z39.19: 2005 standard [21] included the criteria for the maintenance of data organization systems by using automatic control of the thesaurus. This standard was intended for presentation of the contents of different KOS¹³, such as synonyms, taxonomy, thesaurus and other types of controlled vocabularies.

Z39.19 standard presumes that the controlled vocabularies which are commonly used for describing the content and definitions are also used to represent the metadata associated with the content of objects (NISO Z39.19: 2005:12). Unlike other standards used up to this moment, the new standard also implies the use of computerized

¹³ Knowledge Organization System



sources in metadata schemas, such as Dublin Core and the standard network access protocols such as the Z39.50 protocol [22].

IFLA¹⁴ published its report on the management of multilingual thesaurus [23] in 2009 in order to supplement the ISO 5964:1985 standard. It also complements the conditions of the NISO Z39.19:2005 standard. The main provisions contained in the report relate to a method of creating a thesaurus in an asymmetrical form and to the links between different controlled vocabularies. By asymmetric thesaurus is meant multilingual thesaurus, wherein the number of descriptions of the identifiers in each language and method of organization of identifiers are optionally the same and relate to different languages.

British Standard BS8723 (Structured vocabularies for information search) was published between 2005 and 2008. It consists of five parts, in which the latter is also known as DD8723-5 devoted to protocols and formats of interchange for communication. It aims to structuration of dictionaries as whole, describing and comparing different types of KOS, such as classification of schemes, taxonomies, material content scheme, thesaurus, lists of authors and ontology.

Different parts of this British Standard have been canceled after the publication of ISO 25964: 2011 [13], entitled *Information and Documentation. Thesaurus and Interoperability with Other Vocabularies. Thesaurus for Information Retrieval*.

The ISO 25964-2: 2013 standard [14] "Information and documentation - Thesauruses and interoperability with other vocabularies - Part 2: Interoperability with other vocabularies" is the continuation of published in 2011, standard ISO 25964-1: 2011 "Information and documentation - Thesauruses and Interoperability with Other Vocabularies - Part 1: Thesauruses for Information Retrieval".

The new standard is applicable to the thesauruses and other types of dictionaries which are generally used for information retrieval. It describes, compares and contrasts elements and possibilities of these dictionaries, which are relevant when there is the need for interaction (interoperability). The standard provides recommendations on the establishment and maintenance of consistency between multiple

thesauruses or between thesauruses and dictionaries of other types.

The basis of the both parts of ISO 25964 was formed by British Standard BS 8723 "Structured vocabularies for information retrieval. Guide" published in 2005-2007 in the four parts. The following table provides a comparative analysis of international standards and GOSTs.

¹⁴ International Federation of Library Associations and Institutions



Table 1. Comparative Analysis Of International Standards And Gosts

Standards of presentation of thesaurus	The title	Application field	The structure and main requirements to the creation of thesaurus	Distinction from the previous
ISO 5964:1985	" Guidelines for the Establishment and Development of Multilingual Thesauruses»	This standard applies to multilingual information retrieval thesauruses	The standard defines the primitive types of the use of different languages by using data indicating the main questions relating to the semantic equivalence	–
GOST 7.24-2007	“System of Standards on Information, Librarianship and Publishing.. The Information Retrieval Multilingual Thesaurus. Structure and Main Requirements to Creation".	This standard applies to multilingual information retrieval thesauruses	Required components of multilingual information retrieval thesaurus : - prolog; - alphabetic leksiko-semantic indexes including means for indicating the equivalence of descriptors in monolingual versions. Classified, permutation, hierarchical and other indexes and lists of special categories of lexical units, as well as the applications containing additional data on development and use of MIPT into composition of the thesaurus are acceptable in the thesaurus	–
ISO 2788:1986	"Guidelines for Establishment and Development of the Monolingual Thesaurus"	This standard applies to the monolingual information retrieval thesauruses.	The ISO 2788-1986 Standard defines the thesaurus as a set of the terms connected between themselves with the appropriate communications (relations).	–
GOST 7.25-2001	"System of Standards on Information, Librarianship and	This standard applies to the monolingual	Information retrieval thesaurus consists of the	–



	Publishing. The Information Retrieval Multilingual Thesaurus. Structure and Main Requirements to Creation".	information – retrieval thesauruses.	prolog, the main part (the lexico-semantic index) and additional parts (systematic, permutation, hierarchical, etc. indexes and lists of special categories LE)	
ISO 13250:2000/2003	ISO 13250 standard defines the thematic maps.	Thematic maps are a metamodel for definition of categories, association and use by means of semantic links in documents and sections irrespective of the contents. Thematic maps set as the purpose the indexation of the contents of the document in general, and also allow realizing multipurpose indexation for similar type of a web documents.		The information is presented in xml format with explanations to the xmt designations. In the standard of xmt format, key elements and concepts are shown (topics, associations, origin, published themes and field of application). Also TM standard typically includes format of electronic interchange thesauruses, although it is not separately identified for them, because, it is one of its applications.
ANSI/NISO Z39.19:2005	Z39.19:2005 included criteria of maintenance of systems of a data structure by means of automatic control of thesaurus.	This standard was directed to presentation of the maintenance of different KOS such as synonyms, taxonomy, thesaurus and other types of controlled vocabularies. The Z39.19:2005 standard was conceived for application to monolingual thesauruses (NISO Z39.19:2005, point 2.4).	NISO implies that controlled vocabularies are commonly used for the description of the contents by determination of concept for representation of meta data of the subjects.	Unlike other standards used up to this moment, the new standard also implies the use of computerized sources in metadata schemas, such as Dublin Core and the standard network access protocols such as protocol Z39.50
IFLA 2009		This standard applies to multilingual information retrieval	Creations of the thesaurus in the asymmetric form and	IFLA published its report on the management of multilingual thesaurus



		thesauruses	communication between different controlled vocabularies. The asymmetric thesaurus is the multilingual thesaurus in which number of descriptions of identifiers in each language and a method of the organization of identifiers are optionally the same and relate to the different languages.	[23] in 2009 in order to supplement the ISO standard 5964:1985. It also complements the conditions of the NISO Z39.19:2005 standard
ISO 25964-1:2011	ISO 25964-1:2011 "Information and Documentation - Thesauruses and Interoperability with Other Vocabularies - Part 1	It is intended for use in applications for information retrieval. It is applicable to the dictionaries used for retrieval and extraction of information from the information resources of all types irrespective of a type of medium (text, sound, photos and video records, physical or multimedia entities), including knowledge bases and portals, bibliographic databases, text, museum and multimedia collections and objects containing in them. ISO 25964-1:2011 is applicable to monolingual and multilingual thesauruses.	The ISO 25964:2011 standard contains the recommendations for development and maintenance of thesauruses.	The Standard also contains a data model and describes the recommended format for import and export of the thesaurus data.
ISO 25964-2:2013	"Information and Documentation – Thesauruses and Interoperability with Other Vocabularies – Part 2: Interoperability with Other		It is directed to the structured dictionaries in general, characterizing and comparing the different KOS types such	The new standards are applicable to thesauruses and other types of dictionaries which are commonly used for



	Vocabularies.		as classification of diagrams, taxonomy, diagrams of contents of material, the thesaurus, lists of authors and ontology. The software acquires other functions: control of the dictionary, formalization methods necessary for compilation of maps between the dictionaries, and also protocols and formats of interchange	information retrieval. Elements and possibilities of these dictionaries are described, compared and opposed when there is a need for interaction (interoperability). The standard contains recommendations about establishment and maintenance of mutual compliance between several thesauruses or dictionaries of other types.
--	---------------	--	--	---

The ISO 25964:2011 standard appeared as a consequence of the revision of the ISO 2788 and ISO 5964 standards which didn't satisfy the modern requirements of indexation processes, as well as the application of network access procedure 25 years on. The American NISO Z39.19:2005 standard and the British BS5723-1-5:2005-2008 were the most significant additions. NISO generalized the standard for "controlled dictionaries". The British standard established connection between the thesaurus and the very thing that isn't the thesaurus with the controlled dictionary. The biggest difference lies in the transition from terminological structurization to conceptualization, in which equivalents lie between concepts, but not between words.

Due to the format ageing there was a transition from paper to electronic option and increase in functionality. Integration of different resources implies development of the mechanisms of a mapping (creation of maps) or construction, which will allow establishing the international interrelation.

For the first time the ISO 25964-1:2011 standard provided the models (diagram) of data for network interaction. The basis for data schemas consists of the data model of the Z39.50 protocol (Zthes data schema) and recommendations of

SKOS¹⁵ (SKOS data schema) which is based on RDF model representation of concepts.

5. ZTHES SCHEMA

According to common ideology Z39.50, access to any database should be carried out through a single standard pattern data, at which all private entities must be correctly displayed. The schema is called Zthes.

Zthes is designed to work with protocol Z39.50. Note that this schema involves the use of a very limited number of relation types between terms. It was specially designed for compatibility.

Between the terms, in accordance with the recommendations of the standard the following types of links are established:

BT – Broader term: that is, the related term is more general than the current one.

NT – Narrower term: that is, the related term is more specific than the current one.

USE – Use instead: that is, the related term should be used in preference to the current one.

¹⁵ Simple Knowledge Organization System



UF – Use for: that is, the current term should be used in preference to the related one

RT – Related term.

LE – Linguistic equivalent: the current term and the related term are preferred terms representing the same concept – or “sufficiently close” concepts - in different languages.

Links BT and NT, and the USE and UF are mutually inverse. Contact RT, and LE are symmetrical.

In addition, such term is determined in accordance with standard recommendations. There are the following types of terms in the Zthes schema:

- TT - the term of upper level , i.e. the term having

no related terms of

- broader class (terms with the W type of connections);
- NT - not the term of upper level , i.e. a descriptor having the W type connection;
- ND - not a generic term;
- NL - fictitious term, i.e., the term is not used for indexing documents, but is included in the hierarchy to indicate the logical basis of section classes.

The following table (Table. 2) shows the main elements of the Zthes schema:

Table 2. Zthes Schema

field	Description
Element thes	
1.1 dc:title	A name given to the thesaurus
1.2 dc:creator	The person primarily responsible for making the thesaurus.
1.3 dc:subject	The topic of the content of the thesaurus.
1.4 dc:description	Description of the contents of the thesaurus.
1.5 dc:publisher	An entity responsible for making the thesaurus available.
1.6 dc:contributor	An entity responsible for making contributions to the thesaurus.
1.7 dc:date	The date associated with an event in the life cycle of the thesaurus.
1.8 dc:type	The nature or genre of the content of the thesaurus.
1.9 dc:format	The physical aspect or representation of a thesaurus.
1.10 dc:identifier	Id of thesaurus, an unambiguous reference to the resource within a given context.
1.11 dc:source	A reference to a thesaurus from which the present thesaurus is derived.



	1.12 dc:language	A language of the thesaurus.
	1.13 dc:relation	A reference to a related thesaurus.
	1.14 dc:coverage	Localization and limits of applicability of the thesaurus.
	1.15 dc:rights	Information about rights held in and over the thesaurus.
Element term		
	2.1 termId	an opaque string of characters which uniquely identifies the term within the thesaurus
	2.2 termUpdate	updates record.
	2.3 termName	the name of the term in a form which may be displayed to a user or used as a search term in a target database.
	2.4 termQualifier	an additional string which, if supplied, qualifies the term name such that the combination of term and qualifier is unique within the thesaurus.
	2.5 termType	an indication of the type of the term, chosen from the controlled vocabulary described below: TT –top term (terms of type BT relations). PT - Preferred term (also known as a descriptor) ND -Non-descriptor: that is, a non-preferred term. NL - Node label: that is, a dummy term not assigned to documents when indexing, but inserted into the hierarchy to indicate the logical basis on which a category has been divided - for example, by function. Also known as a guide term or a facet indicator.
	2.6 termLanguage	the language of the term
	2.7 termVocabulary	Note is taken from the vocabulary of the term, if the thesaurus contains several dictionaries.
	2.8 termCategory	It defines the term as belonging to a particular subset of the actual (microthesaurus).
	2.9 termStatus	Status term, which may be active, deactivated or deleted. Only active terms may be used for the search.



	2.10 termApproval	Note that, whether the term approved for inclusion in the thesaurus (waiting for authorization or will not be considered for inclusion in it).
	2.11 termSortkey	The key for the sort term.
	2.12 termCreatedDate	the date on which the record defining the term was created
	2.13 termCreatedBy	the name of the person who created the record defining the term.
	2.14 termModifiedDate	the date on which the record defining the term was last modified.
	2.15 termModifiedBy	the name of the person who last modified the record defining the term.
	2.16 termNote	a scope note for the term: that is, arbitrary prose clarifying the meaning and scope of the term.
Element postings		a sub-record, in the format described below, indicating the frequency with which the term occurs in a target database.
	3.1 sourceDb	the host, port and name of a target database in which the term may be found.
	3.2 element fieldName	if specified, the name of a field in the target database in which the term may be found; otherwise, the sub-record represents a postings count across the entire target database.
	3.3 element hitCount	the number of occurrences of the term in the target database (in the nominated field only, if specified).
Element relation		a sub-record, in the format described below, briefly describing a term related to this one.
	4.1 relationType	an indication of the type of the relation, chosen from the controlled vocabulary described below 'NT' - Narrower term: that is, the related term is more specific than the current one. BT - Broader term: that is, the related term is more general than the current one. USE - Use instead: that is, the related term should

		<p>be used in preference to the current one.</p> <p>UF - Use for: that is, the current term should be used in preference to the related one</p> <p>RT - Related term.</p> <p>LE -Linguistic equivalent: the current term and the related term are preferred terms representing the same concept - or "sufficiently close" concepts - in different languages.</p>
--	--	--

6. SKOS SCHEMA

SKOS (Simple Knowledge Organization System) is a subset of the RDF language, used to create a model that expresses the basic structure and content of such concept schemas as thesauruses, classification schemas, lists of named objects, taxonomies and other similar types of dictionaries. As an application of the RDF, SKOS allows you to publish the terms in the web environment to connect them with the information elements and incorporate them into the other conceptual schemas.

SKOS also provides lightweight conceptual modeling language and can be used in combination with more formal languages, such as OWL.

The SKOS main elements:

- Concept - defines the key term, the notion, the idea, the essence of the object domain;
- Semantic relation - correlates the two concepts with one another. SKOS defines two types of semantic relations: hierarchical (Broader / Narrower, BT / NT, Wider / Narrower), and non-hierarchical (Related, RT, associated).

The following table (Table. 3) describes the main elements of the SKOS schema:

Table. 3. SKOS Schema.

field	Description
1. Concepts	
skos:Concept	The name of the concept(column)
skos:ConceptScheme	It defines the concept of scheme
skos:inScheme	To turn the concept into the scheme
skos:hasTopConcept	To specify the root of the concept in the scheme
2. Labels & Notation	
skos:prefLabel	preferred label. There can be only one for each language;
skos:hiddenLabel	hidden label. Used to set the information available for processing, but hidden from the display, for example, for the term of erroneous embodiment;
skos:altLabel	alternative term (label) . It can be used along with the preferred;



3. Documentation		
	skos:scopeNote	Information about the meaning of the concept, if it is or what their limits are;
	skos:definition	complete definition;
	skos:example	includes an example of the spirit, described the concept;
	skos:editorialNote	official comment of the author of the dictionary, thesaurus, descriptive scheme;
	skos:changeNote	official record of the change of the concept or its attributes;
	skos:historyNote	It describes the significant changes in the value or shape concepts;
	skos:note	Notice to clarify the meaning and scope of the concept.
4. Semantic Relations		
	skos:broader	related term has a broader meaning than the current, (the connection to the parent term, i.e. the term more broadly) .
	skos:broaderTransitive	to determine transitive properties
	skos:narrower	related term has a narrower meaning than the current; (communication with the child term , i.e. the term of the narrower sense)
	skos:narrowerTransitive	to determine transitive properties
	skos:related	related term is associative with respect to the current ;
5. Mapping Properties		
	skos:relatedMatch	parallel relations with respect to related concepts from other schemes.
	skos:narrowMatch	analogy relations narrower with respect to the concept of the other schemes;
	skos:exactMatch	the concept can be replaced with a concept associated with it according to the ratio;
	skos:closeMatch	the concept is very similar to the definition, the associated data of attitude;
	skos:broadMatch	parallel relations with respect to broader concepts from other schemes;
6. Collections		
	skos:Collection	collection of terms
	skos:OrderedCollection	It allows defining an ordered set of concepts;
	skos:member	It allows defining a collection of terms ordered;
	skos:memberList	It allows you defining a collection of terms unordered;

Conceptually the SKOS data schema is an equivalent to the Zthes data schema except for the postings element, which is not represented in the SKOS data schema.

7. OTHER SCHEMES

There are the other data schemas for representing thesauruses and controlled vocabularies, but they are all poorer in description of properties and do not meet the standard. The following schemas are worth noting: MARC, VDEX and MODS.

MARC is Machine-Readable Cataloging, which is the machine-readable format of cataloging record [24]. It can only be used for presentation of controlled vocabularies without semantic relations. The MARC format determines the structure and semantics of bibliographic information, especially a standard data structure, rather than the standard content. The content of the recording is subject to terms of cataloging, indexing systems, etc. At the same time there is a number of issues, the interpretation of which is different in the format and cataloging.

VDEX¹⁶ is Vocabulary Definition Exchange, which is the model dictionary for information exchange. The IMS VDEX format is a standard representation of lexicographical resources, regardless of their application, which creates conditions for their portability and sharing by different systems.

MODS¹⁷ is the Metadata Object Description Standard. Driving MODS, developed by the Library of Congress in 2002, is a shortened, more "friendly" to the user version of MARC - a subset of the key elements of data MARC translated into easily understandable XML-based format. Instead of three-digit field labels, abstract subfield identifiers are used which are comprehensible to the user, verbal labels (for example, "title" instead of "245"). Most of the elements of the fixed length data are ignored.

The new identified data elements, such as: "name", which includes the personal name and the name of the organization and can be used in the field of the author, and as a part of the subject headings. Although the MODS circuit is based on MARC21 and is much more detailed than the DC, it has significantly less rules than MARC. As in DC, there are no mandatory fields, all the fields

can be repeated. The MODS entries are often used in databases, which include a mixture of library cataloging and bibliographic data from other sources.

8. IMPLEMENTATION OF THESAURUS

This analysis was carried out in order to choose the approach to implementation of the thesaurus on informatics for supporting "The Modern Problems of Informatics" and "Computing Systems" courses¹⁸. As a platform for implementation of the thesaurus "The Electronic Libraries Control System" was used, developed in the Institute of Computing Engineering, the Siberian Branch of the Russian Academy of Sciences [25, 26].

For implementation of the thesaurus the Zthes data schema was selected, as the most advanced scheme from standard circuits. The principal advantage of the Zthes scheme is its compliance with the model of the network Z39.50 protocol that will allow working not only with personal local thesaurus, but also connecting the thesauruses located in a network if it is necessary. [27].

Three elements were added to the Zthes data schema:

- termNormName is a normal form of the term (singular, the Nominative case, etc.);
- termLinkID is an algorithmically calculated unique identifier, which characterize the term;
- termScopeNote is a short text characteristic of the term, used for identification of the texts.

The controlled vocabulary of relation type (relationType) was expanded to signs of 'SYN' – a full synonym where the full conceptual equivalent of the term received as a rule from distinctions in writing of the term (for example, computing systems ≡ comp. systems).

The created components of implementation of the thesaurus allow viewing, editing and adding terms of the thesaurus to the system through the web forms and also to import and export the term in the form of XML, RDF, DTD files. Data transformation to the SKOS, MARC, MODS data

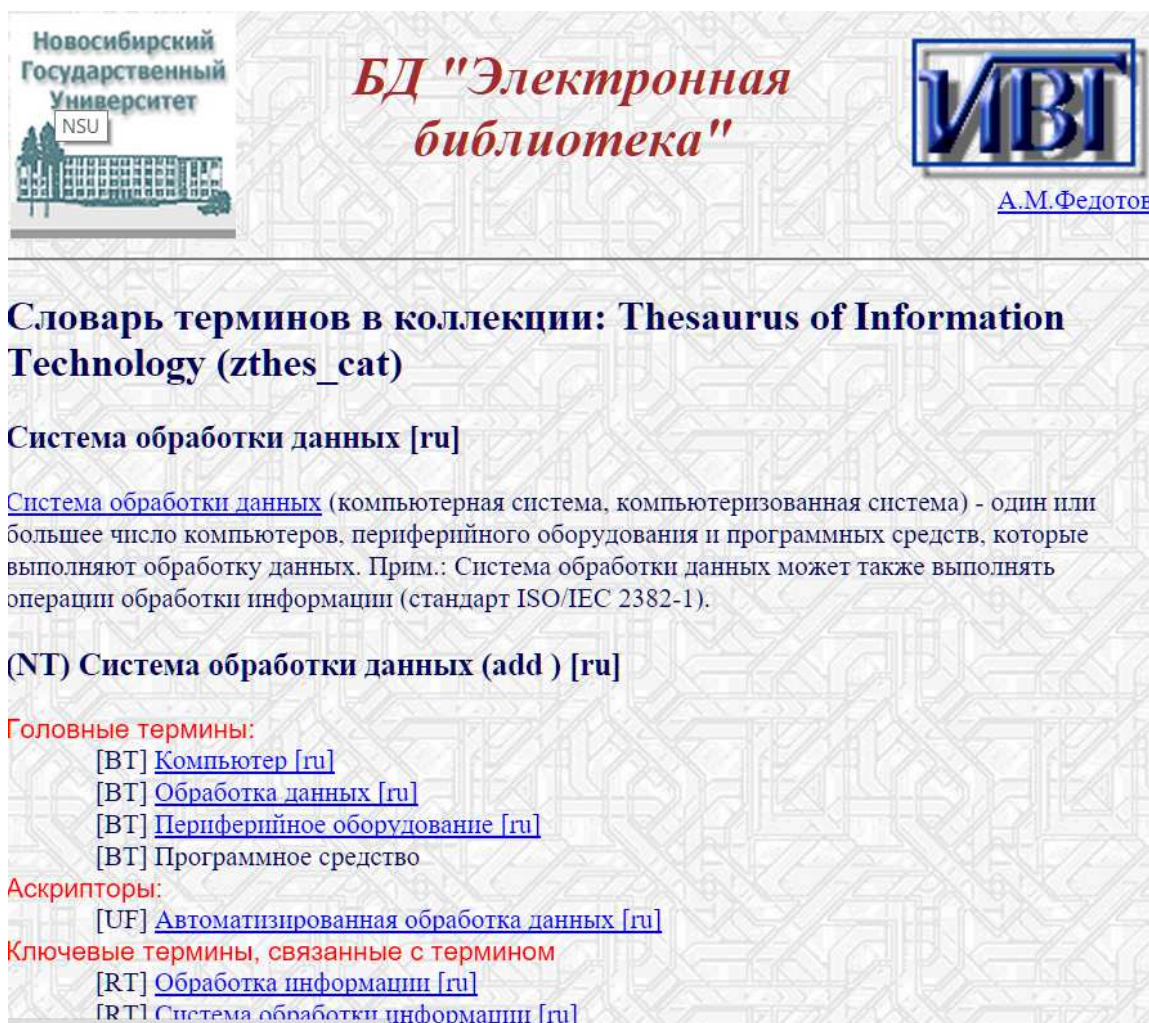
¹⁶ <http://www.imsglobal.org/vdex/>

¹⁷ <http://www.loc.gov/standards/mods/>

¹⁸ <http://fedotov.nsu.ru/infortech/>

schemas and presentation of terms in the browser is made by means of XSLT transformation.

Figure 2 Presentation of a term through the web interface.



Новосибирский Государственный Университет
NSU

БД "Электронная библиотека"

ВБТ

[А.М.Федотов](#)

Словарь терминов в коллекции: Thesaurus of Information Technology (zthes_cat)

Система обработки данных [ru]

[Система обработки данных](#) (компьютерная система, компьютеризованная система) - один или большее число компьютеров, периферийного оборудования и программных средств, которые выполняют обработку данных. Прим.: Система обработки данных может также выполнять операции обработки информации (стандарт ISO/IEC 2382-1).

(NT) Система обработки данных (add) [ru]

Головные термины:

- [BT] [Компьютер \[ru\]](#)
- [BT] [Обработка данных \[ru\]](#)
- [BT] [Периферийное оборудование \[ru\]](#)
- [BT] Программное средство

Аскрипторы:

- [UF] [Автоматизированная обработка данных \[ru\]](#)

Ключевые термины, связанные с термином

- [RT] [Обработка информации \[ru\]](#)
- [RT] Система обработки [информации \[ru\]](#)

Figure 2. Web Presentation of the Term

Распределенная система интеграции данных ZooSPACE

Шлюз Z39.50 - HTTP (Z-GW ZooPARK)

Искать	<input type="text" value="Систем"/>	?	в поле	<input type="text" value="4 - Заглавие"/>	набор	<input type="text" value="Default"/>	с
Количество записей		Поиск SRU		Поиск Z39.50			

Запись: 4 из 24 Представление: **Формат:** **Схема:**

```
<?xml version="1.0" encoding="UTF-8"?>
<Zthes xmlns:dc="http://purl.org/dc/elements/1.1/">
  <thes>
    <dc:title>Тезаурус по информатике</dc:title>
    <dc:creator>Федотов А.М.</dc:creator>
    <dc:rights>ИВТ СО РАН</dc:rights>
    <dc:language>ru</dc:language>
    <dc:language>kz</dc:language>
    <dc:identifier>http://db4.sbras.ru:210/th_compisci</dc:identifier>
    <thesNote>Тезаурус по информатике создан в рамках работ по ...</thesNote>
  </thes>
  <term>
    <termID>1255</termID>
    <termQualifier>D84EEBA6</termQualifier>
    <termName>Система обработки данных</termName>
    <termLanguage>ru</termLanguage>
    <termNote> (компьютерная система, компьютеризованная система) - один или бо
    периферийного оборудования и программных средств, которые выполняют обработку
    данных может также выполнять операции обработки информации (стандарт ISO/IEC
```

Figure 3 The Retrieval Result of the Same Term According to the Z39.50 Protocol

9. CONCLUSION

On the basis of the analysis of standards and different approaches to implementation of thesauruses the decision is made to use the Zthes data schema for creation of the thesaurus on informatics. At this moment the thesaurus contains 1841 terms and it is permanently replenished. Use of the thesaurus in electronic libraries is the most effective in case of its continuous upgrade, the integration into the database and the appropriate level of subject specialization. For the time being the main use of the thesaurus is navigation on the library resources and classification or rubrications.

REFERENCES:

- [1] A.M. Fedotov, O.A. Fedotova, "A model of information system to support scientific and educational activities", *Computational and*
- [2] A.M. Fedotov, O.L. Zhizhimov, O.A. Fedotova, V.B. Barakhnin, "A model of information system to support scientific and educational activities", *Vestnik NSU. Series: Information Technologies*, vol., 12, no. 1, 2014, pp. 89-101.
- [3] V.B. Barakhnin, A.M. Fedotov, "Studying the information needs of scientific community for constructing the information model of its activity", *Vestnik NSU: Information Technologies*, vol. 6, no. 3, 2008, pp. 48-59.
- [4] Yu.I. Shokin, A.M. Fedotov, V.B. Barakhnin, *Problems of information retrieval*, Novosibirsk: Nauka, 2010.
- [5] N.V. Lukashevich, *Thesaurus in the problems of information retrieval*, Moscow: Moscow State University Press, 2011.
- [6] G. Salton, *Dynamic information and library processing*, N.J.: Prentice Hall, 1975.

Informational Technologies in Science, Engineering and Education CIT 2013: Proceedings of the International Conference, vol. 2, 2013, pp. 249-265.



- [7] A.I. Mikhailov, A.I. Chernyi, R.S. Gilyarevskiy, *Scientific communications and informatics*, Moscow: Nauka, 1976.
- [8] O.L. Zhizhimov, A.M. Fedotov, O.A. Fedotova, "Building a generic model of information system for working with documents on the scientific heritage", *Vestnik NSU: Information Technologies*, vol., 10, no. 2, 2012, pp. 5-14.
- [9] ISO 5964:1985. *Guidelines for the establishment and development of multilingual thesauri*. Geneva: International Organization for Standardization, 1985.
- [10] ISO 2788:1986. *Guidelines for the establishment and development of monolingual thesauri*. 2nd ed. Geneva: International Organization for Standardization, 1986.
- [11] M. Masterman, "Semantic message detection for machine translation, using an interlingua", *Proc. 1961 International Conf. on Machine Translation*, pp. 438-475.
- [12] Yu.A. Schrader, "On the quantitative characteristics of semantic information", *NTI Ser.2*, no. 10, 1963, pp. 35-39.
- [13] ISO 25964-1:2011 Information and documentation, Thesauri and interoperability with other vocabularies, Part 1: Thesauri for information retrieval, 2011.
- [14] ISO 25964-2:2013 Information and documentation, Thesauri and interoperability with other vocabularies, Part 2: Interoperability with other vocabularies, 2013.
- [15] GOST 7.25-2001. Thesaurus monolingual information retrieval, Rules for the development, structure, composition and form of presentation (System of standards on information, librarianship and publishing), Interstate Council for Standardization, Metrology and Certification, Moscow: Standartinform, 2002.
- [16] GOST 7.24-2007. Thesaurus multilingual information retrieval, Composition, structure and basic requirements for construction: an interstate standard (System of standards on information, librarianship and publishing), Interstate Council for Standardization, Metrology and Certification, Moscow: Standartinform, 2007.
- [17] V.D. Solovyev, B.V. Dobrov, V.V. Ivanov, N.V. Lukashovich, *Ontologies and thesauri*, MV Lomonosov Moscow State University, 2006.
- [18] ISO/IEC13250:2003. Information technology, SGML applications, Topic maps.
- [19] ISO/IEC 13250-2:2006. Information technology, Topic Maps, Part 2: Data model.
- [20] ISO/IEC 13250-3:2013. Information technology, Topic Maps, Part 3: XML syntax.
- [21] ANSI/NISO. Z39.19: 2005 Guidelines for the construction, format and management of monolingual controlled vocabularies, Bethesda, MD: NISO Press, 2005.
- [22] ANSI/NISO Z39.50-2003. Information Retrieval (Z39.50): Application Service Definition and Protocol Specification, Bethesda, Maryland, USA: NISO Press, November 2002.
- [23] IFLA. Guidelines for multilingual thesauri, IFLA professional (IFLA professional reports: 115), IFLA, 2009.
- [24] RUSMARC. URL: <http://www.rba.ru:8101/rusmarc/>
- [25] Yu.I. Shokin, A.M. Fedotov, O.L. Zhizhimov, O.A. Fedotova, "The control system of electronic libraries", *XV Russian conference with international participation "Distributed information and computational resources"*, DICR-2014: Novosibirsk, December 2-5, 2014.
- [26] Yu.I. Shokin, A.M. Fedotov, O.L. Zhizhimov, O.A. Fedotova, "The control system of digital libraries in IRIS SB RAS", *Infrastructure scientific information resources and systems: Collection of scientific articles of the Fourth All-Russian Symposium*, vol. 1, 2014, pp. 11-39.
- [27] O.L. Zhizhimov, A.M. Fedotov, Yu.I. Shokin, "Technology platform for the mass integration of heterogeneous data", *Vestnik NSU. Series: Information Technologies*, vol. 11, no. 1, 2013, pp. 24-41.