Journal of Theoretical and Applied Information Technology
31st May 2021. Vol.99. No 10

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

EduDIS CONSTRUCTION TECHNOLOGY BASED ON Z39.50 PROTOCOL

[1]S.K.SERIKBAYEVA, [1]D.A.TUSSUPOV, [1,2]M.A.SAMBETBAYEVA, [2]A.S. YERIMBETOVA, [3]TASZHUREKOVA ZH.K., [3]BORANKULOVA G.S.

[1] L.N.Gumilyov Eurasian National University, Department of Information Systems, Astana, Kazakhstan
[2]Institute of Information and Computational Technologies CS MES RK, Almaty, Kazakhstan
[3]Taraz Regional University named after M.KH.Dulaty, Taraz, Kazakhstan

E-mail: [1]inf_8585@mail.ru, [1]tussupov@mail.ru , [1,2]madina_jgtu@mail.ru, [2]aigerian@mail.ru, [3]taszhurekova@mail.ru, [3]b.gau@mail.ru

ABSTRACT

The most effective way to solve the problems of organizing access to distributed information resources is to organize information about them in information systems that are open in the form of electronic libraries. We formulated the basic principles of EduDIS implementation based on the use of the ideas of the electronic library. The article is about the implementation of an information system that provides unified access to distributed resources. The paper considers general issues of organizing access to distributed information resources, as well as the main components of such a system. The main aspects of joint use of Z39.50 protocols for access to distributed information resources are disclosed. In our paper, we provided a functional overview of the protocol itself and the history of standards, described some of the European initiatives that use it, and addressed various questions related to its future use and adoption. Z39.50 is an important building block of future distributed information systems. Moreover, in this paper, we discuss the description of the search for terms in the thesaurus and the algorithm for constructing a tree of terms based on abstract access to the thesaurus, organized in accordance with the Zthes profile of the Z39.50 protocol.

Keywords: *distributed information systems, EduDIS, Z39.50, thesaurus, Zthes, RPN.*

1. INTRODUCTION

In modern society, there is a need to improve the quality and speed of processing, first of all, "big data" [1] and, secondly, data in distributed systems [2]. In this regard, the importance of distributed storage systems [3] and data processing as a means of solving this problem is increasing. One of the main tasks of any distributed system is to analyze the properties of the obtained data, which, for a number of reasons, cannot be evaluated at a single node. To achieve this goal and speed up the processing time, it is necessary at the first stage to send data to distributed nodes of the system, at the second to collect data from distributed nodes and aggregate this data into a general global view. This is a difficult task due to the dynamics that are often encountered in these types of problems, which imposes very frequent changes in local values that affect the general global properties of the entire problem. Creation of efficient and adaptive distributed systems can significantly accelerate the speed of data processing. In order to consider this issue, we will analyze the problems that arise during the design and operation of distributed systems.

For distributed information systems, which include many different databases with different structures and contents, the issue of searching for information in databases using ontologies, thesaurus and classification schemes presented in the form of separate databases is very relevant.

There are many different ways to build a database, organize access to their content and implement explicit and implicit connections between the database and other information resources. Many of these methods are based on strict ontological models and, for practical implementation, impose very strict requirements on the organization of information systems and databases, up to the complete reloading of information into intermediate storages, the functional properties of which make it possible to identify all semantic relationships between information objects based on the given ontological models. This approach has already been discussed in other literature; however, the question remains of how to enable the search for semantically related information in existing distributed information resources, and in the case when they cannot be overloaded into specialized repositories [4].

As a result of using this protocol, it is possible to create distributed information systems, which include databases of various organizations.

In the global network, there are a large number of information resources, each of which has its own form of presentation and storage of information. This causes a global problem of uniform access to information. Each database has a unique structure for storing information, where each field has its own name and purpose.

Also, different rules and programming languages are used to compose search queries to different information resources. In addition, the way of presenting the found information is different. These features cause certain difficulties during the search for information and its issuance. Therefore, each information resource has its own individual interface, which is suitable for storing the data formats used. The user is forced to master a new interface each time and use the information format provided by the resource.

A paradigm is a collection of fundamental scientific knowledge, concepts and terms, accepted and shared by the scientific community. The programming paradigm defines the initial conceptual scheme for setting tasks and how to solve them.

The implementation of the abstract search paradigm today exists in the form of several models for organizing search services, for example, the Z39.50 model and the simpler SRW/SRU model [5]. The pinnacle of excellence in the formulation of attributive search terms belongs to the Z39.50 protocol.

The practical implementation of abstract search services will give a significantly new quality of thesaurus - the ability to include their resources in global search engines at a higher level than the level of external indexing of static WEB pages by other systems [6].

There are many technological developments of information systems for digital libraries. The most famous in Russia are, for example, Euro CRIS (eurocris.org), eLibrary (elibrary.ru), Informika (informika.ru), MathNET (mathnet.ru). In Kazakhstan, the Republican Interuniversity Electronic Library (RI EB) operates on the basis of university educational resources using a unified technology proposed by the Association of Universities of the Republic of Kazakhstan and the scientific and educational computer network of Kazakhstan KAZRENA and the Institute of Information and Computing Technologies of the Committee of Science of the Ministry of Education and Science of the Republic of Kazakhstan. One of the pioneers, and as a result, one of the world leaders in the development of the scientific and educational information system is the "Electronic Library of the Siberian Branch of the RAS", which is now called the Portal of the SB RAS (sbras.ru). This resource, according to Webometrics, ranks first in Russia among scientific organizations and 45th in the world [7].

Educational distributed information systems (Educational distributed information systems - EduDIS) are widely used in the construction of corporate information systems. The concepts of building distributed information systems are characterized by the presence of a large variety of technologies for their implementation.

EduDIS is focused on the creation of distributed information systems (DIS) that integrate heterogeneous information resources managed by various DBMS, based on uniform policies for organizing access to these resources. EduDIS software components operate on various server hardware and software platforms, concentrated in the nodes of a distributed system. The interaction of nodes with each other is carried out by means of network protocols of the application layer based on the transport protocol TCP / IP. The number of nodes in EduDIS is not standardized and can be anything. The EduDIS system can consist of one single node. This choice of node infrastructure allows providing a sufficiently flexible distributed information system and implementing all the necessary functionality that is provided by EduDIS subsystems.

One of the possible options for the implementation of EduDIS is their construction based on the ANSI / NISO Z39.50 protocol. The Z39.50 protocol is intended for communication between computer systems and defines the information retrieval model and the format of its output, but does not define the formats for storing information in databases [8].

Therefore, the task that we have is to develop an application that allows obtaining unified access to a thesaurus with different data structures. Access should be done using specialized protocols that describe the attributes for constructing abstract queries. For the organization of unified access to the thesaurus database, approaches have been developed based on the Z39.50 [9] and SRU/SRW [10] protocols.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Requirements for a distributed information system. Application of the Z39.50 protocol allows organizing a general-purpose distributed information system with rich functionality. Such a system should ideally meet the following requirements:

- Working with distributed data - the information system must provide a possibility to work with data located on different physical servers, different hardware and software platforms and stored in various internal formats.
- Logical grouping of data - the system should allow processing all queries on logical groups of databases, thereby completely hiding the physical location of the latter.
- Abstract data model - an information system is built on the basis of an abstract data schema to which concrete databases must be mapped. This allows data from disparate systems to be combined into one logical grouping.
- Abstract query system - the system should operate not with a specific query syntax, but with its logical essence based on abstract attributes.
- Meta-information - the system must provide complete information about itself and about all its resources.
- Access control - the system should be able to provide different levels of privileges for users to access information.
- Accounting and control - the system should be able to collect statistics on user requests and keep their budgets.
- Openness - the system should be easy to expand and be based on open standards and protocols.
- Communication with other systems - the ability to integrate your resources with the resources of other information systems.
- Ease of communication - for users, the system should provide simple interfaces to access information. A necessary element of the system today is also the presence of a WEB gateway.

Z39.50 can help the library in a number of ways, but there are also problems with the provided capabilities. Some of them are listed below:

Table 1: SWOT analysis

Capabilities	Problems
✓ Provides a single interface for finding information - choose your client carefully ✓ Provides access to many different databases, not just databases online library supports simple and complex searches ✓ Provides flexible and customizable search	✓ With sophistication comes complexity ✓ Z39.50 is misunderstood and invisible to most users ✓ Unrealistic expectations
Advantage	**Disadvantages**
✓ It will be easier for visitors to access information outside the local library. ✓ Access is provided to multiple databases through one interface ✓ Access is provided to several types of data through one interface ✓ You can easily create a "consolidated" directory	✓ Costs - Grant Verification ✓ Complexity: Tuning requires an understanding of both the library database and the server. ✓ Increased visibility and use

The information system must contain the required number of database servers (Z39.50 servers) connected by a network, a WEB server with a WWW-Z39.50 gateway and specialized client workstations. The latter component is necessary for users with increased requirements for the information system; for others, access to it is possible via the WEB (see Fig. 1).

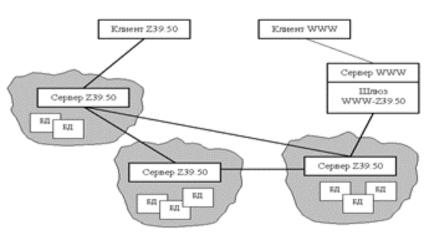

Fig. 1 Information system diagram

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

When discussing technologies for building information systems and distributed databases, the question of processing distributed transactions is usually raised. Undoubtedly, this is the most complex part of the system, allowing remote dynamic updating of data with updating of the latter in real time. However, in the case of a public information system, the task of updating data is separated from the tasks of finding and providing information. The task of updating data is an administrative task that can be solved for each specific DBMS in its own way. As for maintaining data integrity, it can be globally guaranteed provided that an information system is built from completely closed modules - groups of databases, each of which is controlled by its own DBMS, which monitors, as far as possible, the integrity of data in its group.

The authors have built a prototype of such an information system [9]. The fact of filling the described system with bibliographic information does not prohibit filling it with data of a different nature. The central place in the information system is undoubtedly the database server supporting the Z39.50 protocol.

Earlier, in their works, they described in detail the process of abstract access to information resources by O.L. Zhizhimov and N.A. Mazov [1,3,5]. Their works were devoted to the application of the technology of the Z39.50 protocol, which has significant differences from other technologies of network access to databases.

O.L. Zhizhimov in his work "Introduction to Z39.50" [14], which contains a complete list of search attributes and rules for constructing abstract queries.

2. BASICS OF THE Z39.50 STANDARD

The Z39.50 protocol defines the order of interaction between the client and the server, the procedures for searching and retrieving information from databases, and the formats for presenting this information.

The Z39.50 protocol does not define the formats for storing data in specific databases, methods of indexing them, and procedures for the functioning of various DBMSs. The Z39.50 protocol also does not define the user-client interfaces.

In the Z39.50 ideologies, within the same schema, all databases are exactly the same, despite their physical differences in the used DBMS, fields and query syntax. In Z39.50, there is no way for the client to determine which database management system is used to store the data it retrieves. This seemingly limitation makes a lot of sense, the essence of which will be clarified in the following sections. However, it can be noted that the client does not need such information, because he always works with the same query system and receives data in the same formats [11].

Information Retrieval (Z39.50): Application Service Definition and Protocol Specification defines an application service and protocol specification for retrieving and retrieving information from databases.

It is designed to unify network access to databases and defines procedures for searching, retrieving, and information presentation formats [12].

The first version of the protocol, called Z39.50, was prepared by the National Information Standards Organization (NISO) and was introduced in 1988 by the Z39.50-1988 standard. The effect of this version was extended only to work with bibliographic information.

In 1989, the Maintenance Agency Z39.50 was organized under the administration of the US Congressional Laboratory, and in 1990 the Implementation Team was formed.

Z39.50 (Z39.50 Implementors Group - ZIG). Producers, sellers, distributors of various types of information (bibliographic, textual, graphic, financial, chemical, etc.) have become its members. The Z39.50 Support Agency is a permanent body that maintains and develops this standard. The agency's network address is https://www.loc.gov/z3950/agency/ here you can find all the information on the protocol, news, documentation, object registers, etc. [13].

In 1992, these organizations developed version 2 of the standard (Z39.50-1992), replacing the 1988 standard. Version 3 of the standard (Z39.50-1995) was developed in 1995. Since the Z39.50-1995 standard is a development of the version 1992 and is compatible with it, it defines the protocol versions 2 and 3. In 1995, the Z39.50 protocol was adopted as the American national ANSI standard (ANSI/NISO Z39.50-1995), and in November 1998 as an international ISO-23950 standard. The current ANSI/NISO Z39.50-2003 [6] standard was approved in November 2002. It is a technical revision of the ANSI/NISO Z39.50-1995 standard and also defines versions 2 and 3, but additionally includes various clarifications, fixes,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

and conventions recommended by the Z39.50 - ZIG Implementation Team.

In the first years of its existence, the protocol was used mainly for organizing access to bibliographic resources; today the scope of its application has been significantly expanded. It is used to access scientific, technical, financial information, geoinformation resources, global metadata databases, thesaurus and rubricators, digital collections [14] and museum information [15].

The Z39.50 standard defines a client / server service and protocol for information retrieval. It specifies procedures and formats for a client to search databases provided by a server, retrieve records from databases, and perform other information retrieval functions. The Z39.50 protocol only defines the interaction between client and server retrieval applications, it does not define the interaction between the client and the end user. More precisely, the Z39.50 protocol does not define the formats for storing data in specific databases, methods of indexing them, and procedures for the functioning of various DBMSs. It also does not define the user-client interfaces.

Without going into the details of the protocol, we can say that the Z39.50 standard defines such rules for the interaction of computers that allow unifying access to various databases. Thus, a user using only one client application can search for information in remote distributed databases with a wide variety of structures and information presentation formats [16].

Two main features distinguish the Z39.50 protocol from other protocols. First, it is an abstract model for presenting the information. In the Z39.50 ideologies, within the same data schema, all databases are exactly the same, despite their physical differences in the used DBMS, fields and query syntax. In other words, the protocol provides an abstract model for representing information at each stage of client-server interaction. In Z39.50, the client always works with the same query system and receives data in the same formats.

The second feature is that the Z39.50 protocol fully provides session communication between the client and the server. This feature is embedded in the protocol itself and is implemented in all its applications, be it a server system or a client program [17].

It should be noted that Z39.50 is not a search engine. The Z39.50 client can send search results to one or more databases on remote systems at the same time. The model allows clients to connect to each individual server, searching for the current contents of the database and retrieving results directly from the original databases [18].

But the web search engine is essentially a single information retrieval system that has the additional function of collecting resources from the Internet and performing a kind of indexing to make those resources searchable. In distributed and integrated access, finding information from a single server is not too difficult, but becomes more problematic if the search occurs from multiple databases on multiple servers. It is also difficult for Z39.50 to understand the requests and responses of the two systems. The lack of semantic interoperability has led to a loss of user confidence in the Z39.50 interface for search engines.

3. THESAURUS IN Z39.50 TECHNOLOGY

As defined by the International Organization for Standardization (ISO), a thesaurus is an indexing language-driven vocabulary formally organized to establish explicit a priori relationships between concepts [19]. This definition establishes lexical units and semantic relations between these units as the elements that make up the thesaurus. Thesaurus relations (genus - species, part-whole, etc.) are imposed on the taxonomy structure, i.e. the main taxonomies of the subject area are identified.

Historically, the thesaurus was created for manual indexing of documents and did not take into account the issues associated with automatic indexing when creating them.

The difficulty of constructing a thesaurus corresponding to the entire thematic variety of indexed information is the main reason for its unpopularity in modern information systems. Nevertheless, the effectiveness of information retrieval systems to support scientific and educational activities directly depends on the use of a specialized thesaurus.

An information retrieval thesaurus (in accordance with the definitions of standards) is a normative (controlled) dictionary of key terms in natural language with explicitly specified semantic relations between terms, designed to describe the content of documents and search queries [20]. The thesaurus is intended to describe a specific subject area, each term of which denotes or describes a concept from this subject area.

Based on the general ideology of Z39.50, access to any database should be carried out through a single standard data schema, to which all private structures should be correctly mapped. The scheme was named Zthes.

Zthes is designed to work using the Z39.50 protocol. Note that this scheme assumes the use of a very limited set of types of relations between terms. This is intentional for better compatibility.

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

• BT - link to the parent term, ie. with a term of broader meaning;

• NT - a relationship with a child term, that is, with a term of a narrower meaning. The BT↔NT relationship is reciprocal;

• USE - link to the term that is used instead;

• UF - mutual feedback USE, USE↔UF;

• RT is a link that defines a related term;

• LE - relationship between linguistically equivalent terms.

Links BT and NT, as well as USE and UF, are reciprocal. The RT, as well as LE bonds, are symmetrical.

In addition, the type of term is determined in accordance with the recommendations of the standard.

The following types of terms are distinguished in the Zthes schema:

• TT is a top-level term, ie. a term that does not have related terms of a broader class (terms with the BT link type);

• NT is not a top-level term, that is, a descriptor with BT-type links;

• ND is a minor term;

• NL is a dummy term, that is, a term not used for indexing documents, but included in a hierarchy to indicate the logical basis of a class section [21].

The thesaurus on IT technologies used in the IS for support of scientific and educational activities is built in a client-server architecture based on the Z39.50 protocol. Z39.50 protocol. defines the mechanism of information exchange in the process of processing search queries and the protocol for the exchange of data in the systems performing the search. Proceeding from the general ideology of Z39.50 [22], access to any database, including the thesaurus database, must be carried out through a single standard data schema, on which all private structures of the thesaurus database must be correctly mapped. In addition to describing the abstract structure of a record, the Zthes profile defines the types of terms and the types of relationships between terms. The authors in [23] developed a thesaurus on IT technologies taking into account the morphology of the Kazakh language.

In connection with the rapid growth of various sources of information, the problem of extracting the necessary data from a huge number of documents is becoming increasingly important. Often, the question is no longer about finding a book or article where information of interest to a person is mentioned, but about choosing the most appropriate and most fully illuminating documents from hundreds and thousands of similar sources. Information retrieval systems can solve this problem.

Information search is a line of research that studies the issues of document search, processing of search results, as well as a number of related issues: modeling, classification, clustering and filtering of documents, design of search engine architectures and user interfaces, query languages, etc. [24].

Any system operating on the Z39.50 protocol must necessarily support the Bib-1 attribute set, which determines how terms are specified in a search query. There are six attribute groups in Bib-1:

a set of Use - the fields themselves, or as they are called "access points";

set Relation - attributes of relations (for example,>, <, =, etc.);

set Position - attributes that determine the location of the search engine

term in a field (first word in a field, in a subfield, any position)

set Structure - attributes defining the structure of the search term (word, phrase, key, etc.)

Truncation set - attributes that determine the possibility of term truncation;

Completeness set - Attributes defining the completeness of a field or subfield.

On the one hand, such granularity in defining the properties of search terms implies quite a lot of flexibility in composing queries, but, on the other hand, it creates obstacles for searching for a single query that is simultaneously addressed to several servers.

This is due to the fact that due to the use of different database management tools on

Journal of Theoretical and Applied Information Technology
31st May 2021. Vol.99. No 10

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

different servers, it is almost impossible to implement all the sets of attributes offered by the standard. It often happens that for a search on one server, specifying an attribute is a prerequisite for a successful search, while on another server this attribute, or even the entire group of attributes, is simply not implemented. In this case, it will not be possible to use a single query to access the databases.

This problem is especially relevant in the case of a search when the exact location of the resource is unknown, and the search must be carried out on several Z servers. First of all, when forming a query, the problem of choosing search fields (access points) to be included in the search expression arises. The sets of access points supported by different Z-servers are often different. Although fields such as Author, Title, or Date of Issue are supported in almost all databases, this is often not enough. And fields such as ISSN, for example, may not be supported by the server. All this entails the unpredictability of search success.

Every system based on the Z39.50 protocol must support at a minimum the processing of requests in RPN (Reverse Polish Notation) format, which is expressed using a set of abstract identifiers (OID numeric indexes). The Z39.50 server, upon receiving such a request, MUST convert it to a form suitable for the specific database. The Z39.50 server can process queries in the form of ZSQL, which can also rely on abstract data structures (for example, instead of the name of a table field, use its numerical value from the corresponding set of attributes, and specify the data schema identifier in the FROM section, etc.) [25]. ZSQL queries can be expressed in the form of "familiar" SQL, but this contradicts the Z39.50 ideology, although it can be used in a narrow circle as an alternative to ODBC, JDBC, etc. technologies, providing a single network interface for data exchange. When using relational DBMS, it is necessary to determine which tables will participate in the query, what the type of relationships between them will be, and take into account the possibility of arbitrary nesting of queries. Figure 1 shows an example RPN and ZSQL query using Bib-1 search attribute OIDs and tags (tagset-G). For more information on requests, see the description of the protocol specifications.

EduDIS uses static configuration files to establish correspondences between real tables and search attributes. These files, which can be unique for each database, must contain information about the correspondence of OID elements to SQL query elements. If we restrict ourselves only to search queries (select), then the configuration files should contain information on the correspondence of table fields to search attributes and data schema elements, on the types of fields, and on the nature of relationships between tables. In the general case, if unified query converters are created, then inevitably there is a need for information about the SQL dialects of a particular DBMS, because, despite the existence of SQL standards, there are some differences in query languages.

In general, the structure of the configuration file is quite complex. But for a number of applications, such files can be significantly simplified. Indeed, very often the volume and nature of information provided by open information systems does not imply the use of complex queries to many related tables. It is often wiser to use stored procedures and pseudo-tables (views) instead of loading the Z39.50 server [26].

a) **@and** **@attr** 1=4 **@attr** 5=3 {system} @attr 1=1005 { OIGGM } @attr 1=1003 {Serikbayeva}

b) **select** [(2,1)]
from [1.2.840.10003.13.2]
where [(2,4)] = (select min([(2,4)])) from [1.2.840.10003.13.2]

c)

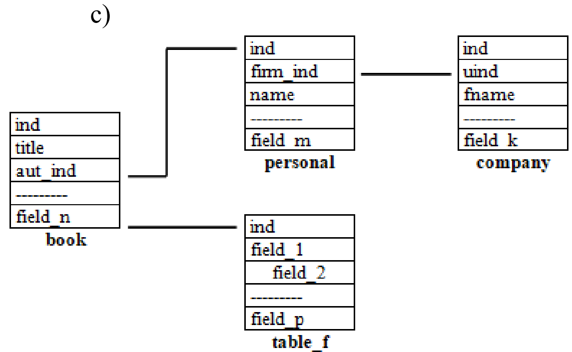

Figure 2. Conversion of requests: a) request in the form of RPN; b) a request in the form of ZSQL-abstract; c) an example of tables and relationships between them in a relational database that contains the required information

To build RPN (PQF) queries, the following model should be used:

query ::= top-set query-struct

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

top-set ::= ['@attrset' attrsetname]
attrsetname :: 'Bib-1' | 'XD-1' | 'util' | 'Zthes'
query-struct ::= simple | complex
complex ::= operator query-struct query-struct.
operator ::= '@and' | '@or' | '@not'
simple ::= attr-spec term.
attr-spec ::= '@attr' [attrsetname] typ '=' value [attr-spec]
typ :: '1' | . . . |'12'
value :: numeric or string
term ::= string.

An example of a simple RPN request is a search request for the term "Information System" that appears in the header (termName):

@attr XD-1 1=1 Information System

where

@attr XD-1 1=1 – corresponds to the "Term name" field;

Information system - search term.

An example of a more complex query to search for the terms "Information system" or "Informatics" found in the headings:

@or @attr XD-1 1 = 1 Information system @attr XD-1 1 = 1 Informatics

An example of a query that finds records that contain "Information system" in the termName field and "ru" in the termLanguage field or " Informatics" in the termName field and "ru" in the termLanguage field:

@or @and @attr XD-1 1 = 1 Information system @attr util 1 = 3 ru

@and @attr XD-1 1 = 1 Informatics @attr util 1 = 3 ru

An example of a query using an additional attribute that allows you to specify a sequence of characters contained in a word as a search term:

@attr XD-1 1 = 1 @attr util 5 = 4 @attr util 9 = 2 {search inf}

The result of such a request may be the terms: "Information search" and "Information search".

An example of a query where a phrase is specified as a search term. A phrase is a set of words separated by a space. The result of the query depends on the sequence of words in the query. In this case, the answer will contain the phrase - "Information search".

@attr XD-1 1 = 1 @attr util 5 = 4 {search inf}

An example of a query that finds all records for the 2019th year, the search is performed by the termCreatedDate field:

@attr util 1=1 @attr util 12=2 {2019}

4. SEARCHING RESULTS

As a result of performing a database search, the client can receive the following information from the server: an error message, the number of records found, or the records found themselves. The first answer is associated with some kind of error, the second and third answers correspond to a successful search. Which one will be received by the client depends on the parameters that are passed to the server along with the search request.

For this, the concepts of small, medium and large sets are introduced. Here, a set is understood as a collection of found records, numbered through. All records from the small set are always returned, all records from the large set are never returned, and some records from the middle set are returned. Next, the following parameters are set.

• The upper limit of the small set, i.e., the maximum record number in the small set that starts with the first record.

• The lower bound of a large set, i.e., the number from which the entries are added to the large set. All records with numbers greater than the upper limit of the small set but less than the lower limit of the large set is considered to be from the middle set.

• The number of records returned from the middle set.

By changing these three parameters, you can achieve the return of any number of records, including none.

Finally, it should be noted that all records found during the search should be saved by the server in the session block for later use. If the server allows the option of naming the search result, this stored collection can be named, if not, the collection is kept unnamed and overwritten in subsequent searches. Named result sets that are stored on the server can be used in subsequent RPN requests, where they act as operands like APT blocks.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

To access the thesaurus database, the user should build an abstract query. For this, a custom WEB application was developed that generates abstract queries to the thesaurus database. To get an abstract query, the user must fill in the form input fields with the following search parameters: the name of the search term, the name of the set of attributes, the search attribute of the Use (Access Point) type. If desired, the user can build queries to find more precise information using additional search attributes and logical operators.

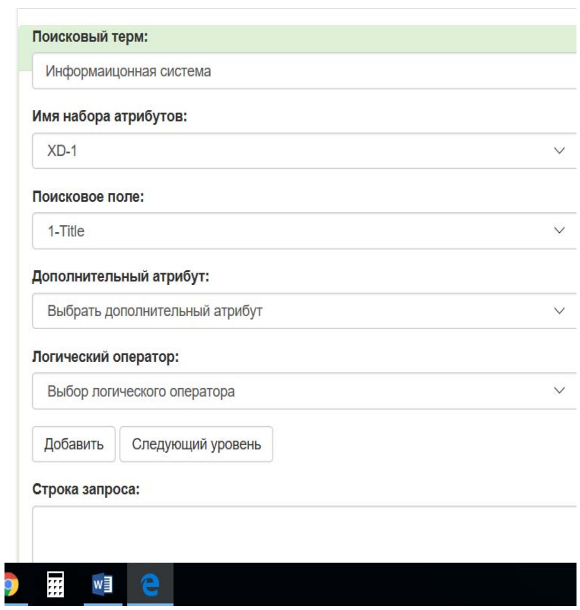

Figure 3. Choice of parameters of an abstract query

After filling in the data, you must click on the "Add" button. The selected search parameters will be displayed in the input field below.

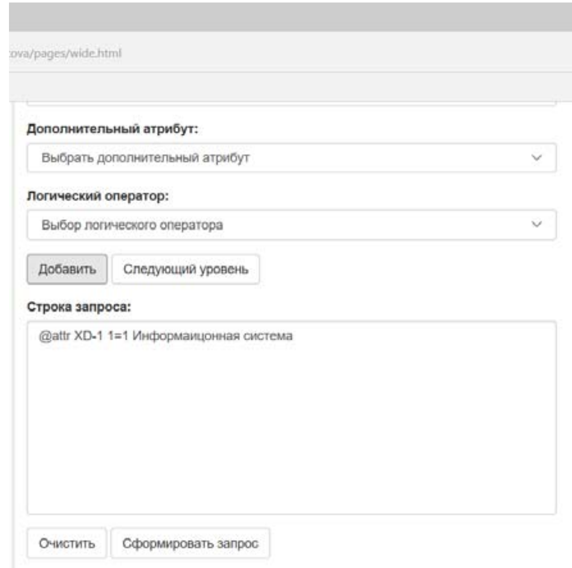

Figure 4. Choice of parameters of an abstract query

To transform an abstract query into a real SQL query to the thesaurus database, a function was developed that is embedded in the DBMS, which works according to the following algorithm.

To begin with, the user must build an abstract query, for example, using the application and check its correctness using the application:

@or @attr XD-1 1 = 1 Information system @attr XD-1 1 = 1 @attr 5 = 1 Informatics

This request in the form of a string is passed to the function as an input parameter. Next, the request is parsed:

@or - logical OR operator
@attr XD-1 1 = 1 - search by "Term name"
Information system - search term
@attr XD-1 1 = 1 - search by "Term name"
@attr 5 = 1 - right truncation
Informatics- search term

The resulting result is replaced with fragments of the SQL query:

@or -> or
@attr XD-1 1 = 1 -> title
Information system -> ‘Information system’
@attr XD-1 1 = 1 -> search by title
@attr 5 = 1 -> LIKE
Informatics -> ‘Informatics%’

Journal of Theoretical and Applied Information Technology
31st May 2021. Vol.99. No 10

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Next comes the arrangement of logical operators and brackets:

((title = 'Information System') or (title LIKE 'Informatics%'))

After that, an SQL query is executed against the thesaurus database:
SELECT * FROM zthes_cat WHERE ((title = 'Information System') or (title LIKE 'Informatics%'))
The resulting result is displayed in a table in the user application:

№	title	link_id	term_qualifier	term_vocabulary	description	document_language	term_category
7	Информатика	54F38E0C	publ555		*Информатика — это наука, которая занимается вычислением, хранением и обработкой информации. Она развивается вместе с компьютерами и сетью интернет, а потому базируется на компьютерной технике и невозможна без нее.*	ru	
2	Информационная система	BFC88BB8	abacus_ru	-	Информационная система — это взаимосвязанная совокупность средств, методов и персонала, используемых для хранения, обработки и выдачи информации для достижения цели управления.	ru	-

Figure 5. Search result

5. CONCLUSION

In our paper, we discussed the basic concepts of a distributed information system and methods for building EduDIS technology in the Z39.50 protocol. We presented the full cycle of work with abstract RPN requests (PQF) of the Z39.50 protocol (from request to execution), and also considered an alternative approach based on SQL query language.

As a result of the work, a function was developed in the built-in language PostgreSQL. The aim of this function is to transform an abstract query into a real SQL query to the thesaurus database.

The function input is a string with an abstract request. First, the syntax of the query is checked, then the search attributes are replaced with real fragments of the SQL query. At the end of the function, the brackets and logical operators are placed.

We also created a user interface written in the PHP programming language. Using the resulting interface, one can generate abstract requests and check their syntax. The user can choose the type of the generated request: the request can be simple or complex [9].

Abstract queries are generated based on search attributes from the Zthes-1, bib-1, XD-1, and util sets. These attribute sets are included in the Use attribute group [10] and are intended to be searched by the thesaurus.

Simple queries are built without using boolean operators and search for one search parameter. Complex queries include logical operators that link multiple simple queries together. To obtain a more accurate search result, when building a query, five groups of additional attributes (Relation, Position, Structure, Truncation, Completeness) are used. Queries can be simple or complex. A recorded WEB application that interactively creates an RPN request, validates it, and then executes it.

Z39. 50 is a client-server architecture standard in which the search engine and interface are separated into independent parts. If both the client and server conform to the standard, the Z39.50 client can search for any brand of Z39.50 server. The most common databases on different local systems can be found through the same local client or through an interface. This does not solve the problem of how the interface should look or how it should act, it is up to the user to choose the interface. The connection of library systems with the Internet and the development of the Z39.50 protocol open up the prospect of access to an ever-growing array of bibliographic databases and full-text databases through a local automated system. The ability to directly link users to resources that provide various computing platforms has made the Z39.50 protocol more attractive to libraries in the institutional systems. As a result of using this protocol, it is possible to create distributed information systems containing databases of various organizations.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ACKNOWLEDGMENT.

«This research has been funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant No. AP09057872)»

REFERENCES:

[1] N.A. Mazov, O.L. Zhizhimov Application of the Z39.50 protocol for working with thesaurus and classification schemes // VII International Conference "Libraries and Associations in a Changing World: New Technologies and New Forms of Cooperation" - Crimea-2000 (Sudak, Ukraine, 03.06 - 13.06.2000): Proceedings of the conference - M.: State Public Library for Science and Technology of Russia, 2000. - Vol. 1. - pp. 402-405.

[2] Yu.I. Shokin, A.M. Fedotov Support and development of distributed information and computing resources of the SB RAS // Computational technologies. (Joint release). Vestnik KazNU them. Al-Farabi. Series: Mathematics, Mechanics, Informatics. Part 4. - 2004. - T.42, Part 4, 3. pp.324-334

[3] O.L. Zhizhimov, N.A. Mazov Model of a distributed information system of the Siberian Branch of the Russian Academy of Sciences based on the Z39.50 protocol. Electronic Libraries, 1999, v.2, issue 2, ISSN 1234-5678.

[4] O.L. Zhizhimov, N.A. Mazov, A.S. Frolov Access to ISIS databases from the Internet and building a distributed information system. Computational technologies, vol. 2, No. 3, 1997, pp. 45-50.

[5] O. L. Zhizhimov, N. A. Mazov, A. Yu. Bolvanov Experience of building a distributed information system based on the Z39.50 protocol // Libraries and associations in a changing world: new technologies and new forms of cooperation: 6th international conf. "Crimea 99" (June 5-13, 1999, Sudak): Materials of the conference, T. 1. - Simferopol: Tavrida, 1999. - pp. 249–252.

[6] O.L. Zhizhimov, N.A. Mazov The current state of the Z39.50 software in the Siberian Branch of the Russian Academy of Sciences // Libraries and associations in a changing world: new technologies and new forms of cooperation: 7th Intern. conf. "Crimea 2000" (June 3-11, 2000, Sudak): Proceedings of the conference, T. 2. - Simferopol: Tavrida, 2000. - pp. 44-46.

[7] Yu.I. Shokin, N.M. Temirbek, O.L. Zhizhimov, A.N. Temirbekov, D.R. Baigereev Model of an integrated distributed library information system of the Akademgorodok of Almaty // Computational technologies. 2018. Vol. 23, No. 5, pp. 110–119. DOI: 10.25743 / ICT.2018.23.5.009.

[8] O.L. Zhizhimov, N.A. Mazov Software Z39.50 - ZooPARK and Z-IRBIS - version 4: new opportunities // Libraries and information resources in the modern world of science, culture, education and business: 11th Intern. conf. "Crimea 2004" (June 5-13, 2004, Sudak): Proceedings of the conf. - M .: Publishing house of the State Public Library for Science and Technology of Russia, 2004. - pp. 702–711.

[9] ANSI/NISO Z39.50-2003. Information Retrieval (Z39.50): Application Service Definition and Protocol Specification // NISO Press, Bethesda, Maryland, U.S.A. — ISBN 1-880124-55-6. — 267 p.

[10] SRU / SRW profile [Electronic resource] URL: zthes.z3950.org

[11] Plemnek A.I. Using the Z39.50 and http protocols in modern library information systems / A.I. Plemnek, R.T. Usmanov, D.N. Owl // Inf. bul. RBA. 1998. - No. 12. - pp. 240-257.

[12] N.A. Mazov, O.L. Zhizhimov, S.V. Skibin Experience of building distributed information systems based on the Z39.50 protocol // Scientific and technical libraries. 2003. - No. 2. - pp. 24-28.

[13] ANSI/NISO Z39.50-1995. Information Retrieval (Z39.50): Application Service Definition and Protocol Specification. Z39.50 Maintenance Agency Offical Text for Z39.50-1995, July 1995.

[14] O.L .Zhizhimov Introduction to Z39.50 // Novosibirsk. Publishing house NGONB. – 2000, 196 p. - ISBN 5-88742-037-5.

[15] O.L. Zhizhimov Introduction to Z39.50: ed. 2nd add. and revised // Novosibirsk. Publishing house NGONB. - 2001 .-- 220 p. - ISBN 5-88742-046-4.

[16] O.L. Zhizhimov, N.A. Mazov Z39.50 in the information community of Russia: state and prospects // Sat. scientific. Art .: Library computer networks: Russia and the West. Issue 2. - M .: Liberia. - 2003. - pp. 99-109.

[17] O.L. Zhizhimov, N.A. Mazov Principles of building distributed information systems based on the Z39.50 protocol. Novosibirsk, 2004.361 p.

[18] O.L. Zhizhimov, N.A. Mazov Model of a distributed information system of the Siberian Branch of the Russian Academy of Sciences based on the Z39.50 protocol. Electronic Libraries, 1999, v.2, issue 2, ISSN 1234-5678.

[19] ISO 25964-1:2011 Information and documentation – Thesaurus and interoperability with other vocabularies – Part 1: Thesaurus for information retrieval, 2011.

[20] A.M. Fedotov, I.A. Idrisova, M.A. Sambetbaeva, O.A. Fedotova The use of the thesaurus in the scientific and educational information system // Bulletin of the Novosibirsk State University. Series: Information Technology. - 2015. - T.13. - No. 2. - P.86-102. - ISSN 1818-7900. - EISSN 2410-0420.

[21] Sambetbayeva M.A., Fedotova O.A., Fedotov A.M. Multilingual Thesaurus in Information System for Scientific and Educational Activity Support // Proceedings of the XX International Conference “Data Analytics and Management in Data Intensive Domains” (DAMDID/RCDL’2018). – M.: Lomonosov Moscow State University, 2018. – pp. 360-362.

[22] ANSI/NISO. Z39.19: 2005 Guidelines for the construction, format and management of monolingual controlled vocabularies. – NISO Press: Bethesda. – 2005, 135 p.

[23] Fedotov A.M., Tusupov J.A., Sambetbayeva M.A., Sagnayeva S.K., Bapanov A.A., Nurgulzhanova A.N., Yerimbetova A.S. Using the thesaurus to develop it inquiry systems // Journal of Theoretical and Applied Information Technology. - 2016. - Vol.86. - Issue 1. - pp.44-61.

[24] R. V. Sharapov, E. V. Sharapova, O. A. Saratovtseva Models of information retrieval. URL: http://vuz.exponenta.ru/PDF/FOTO/kaz/Articles/sharapov1.pdf

[25] S.A. Santeyeva, O.L. Zhizhimov User interfaces for working with thesaurus and rubricators in distributed heterogeneous information systems on the example platform ZooSPACE // Joint issue based on the materials of the international scientific conference "Computational and Information Technologies in Science, Technology and Education" (CITech-2015) (September 24-27, 2015): Computational Technologies, v.20; Bulletin of KazNU them. Al-Farabi, Series of mathematics, mechanics and informatia № 3 (86) / KazNU im. al-Farabi. - 2015. - Part I. - Almaty - Novosibirsk. - pp.224-230.

[26] O.L. Zhizhimov, Yu. V. Titova, A.M. Fedotov, “Implementation of methods for abstract access to a thesaurus”, Vestnik NSU. Series: Information Technology, 15: 1 (2017).