DEVELOPMENT OF A MODEL AND TECHNOLOGY OF ACCESS TO DOCUMENTS IN SCIENTIFIC AND EDUCATIONAL ACTIVITIES

The paper deals with general issues of organizing access to electronic documents in the framework of scientific and educational activities.

Large volumes of already existing information, its continuous growth, the heterogeneous nature of storage and distribution, the lack of a unified way of working with it create many difficulties when using it. Awareness of these difficulties, qualitative changes in the field of information technology and telecommunications have led to the need to solve the problem of finding new approaches to the creation of repositories of information resources, their structure, and the development of tools necessary for users. Currently, such approaches are called "digital" or "electronic" libraries.

According to the preliminary concept, an intelligent scientific and educational Internet resource will be an information system accessible via the Internet, providing systematization and integration of scientific knowledge, data, and information resources into a single information space, meaningful and effective access to them, as well as support for their use in solving various scientific and educational tasks.

Another problem of the organization of effective information support for scientific and educational activities is that, due to its diversity and multidimensional nature, scientific and educational information resources are dispersed on remote pages of many sites and in distributed electronic libraries and archives. To solve this problem, it is necessary to solve the problem of bringing such resources related to one area of knowledge into a single information space, and also, no less important, to support their logical integrity. Without solving these two related tasks, it is impossible to solve the main task – to provide all participants of scientific and educational activities with meaningful access to integrated information resources and means of their analysis.

The support of information systems in the field of scientific and educational activities is relevant, since the need for information always exists. In order to satisfy this need, it is necessary to organize access to various resources

Keywords: information systems, information resources, access, document, distributed, metadata

S a n d u g a s h   S e r i k b a y e v a
Corresponding author
Doctoral Student, Teacher*
E-mail: inf_8585@mail.ru
J a m a l b e k   T u s s u p o v
Doctor of Physical and Mathematical Sciences, Professor, Head of Department*
M a d i n a   S a m b e t b a y e v a***
PhD, Associate Professor, Leading Researcher*
A i g e r i m   Y e r i m b e t o v a
PhD, Associate Professor, Leading Researcher***
Department of Software Engineering
Institute of Automation and Information Technologies Satbayev University
Satbayev str., 22a, Almaty, Republic of Kazakhstan, 050013
Z h a n n a   S a d i r m e k o v a
Postgraduate Student, Master of Technical Sciences
Department of Physical and Technical Informatics
Novosibirsk State University
Pirogova str., 1, Novosibirsk, Russian Federation, 630090
A i g u l   T u n g a t a r o v a
Candidate of Pedagogical Sciences
Department of Information Systems
M. Kh. Dulaty Taraz Regional University
Tole bi str., 60, Taraz, Republic of Kazakhstan, 080000
A r d a k   B a t y r k h a n o v
Master, Senior Lecturer
Department of Software Engineering
Kh. Dosmukhamedov Atyrau University
Studenchesky ave., 1, Atyrau, Republic of Kazakhstan, 060011
A l m a   Z a k i r o v a
Associate Professor, Head of Department
Department of Science and Certification of Scientific Personnel**
*Department of Information Systems**
**L. N. Gumilyov Eurasian National University
Satpayev str., 2, Nur-Sultan, Republic of Kazakhstan, 010000
***Institute of Information and Computational Technologies
Shevchenko str., 28, Almaty, Republic of Kazakhstan, 050010

1. Introduction

An important stage in the informatization of society is the gradual transition from classical paper documentation to electronic documents. The use of automated documentation management systems is of great importance in the modern information world.

Recently, users of traditional libraries have faced a number of problems when accessing library facilities, some of which include insufficient access time, distance barrier, bulkiness, large volumes and heavy weight of some printed books and other materials, which makes them inconvenient for mobility. In addition, library administrators and librarians face difficulties in managing the library premises in

organizing the cataloging of books and replacing used materials on their various shelves and materials, since there are often cases of loss or incorrect placement.

Document access technology based on the specific work of a research organization should meet the following basic requirements: interoperability based on open systems standards, integration into a unified information environment, distribution of documents across storage locations and decentralization of document collection administration, use of off-the-shelf software solutions, provision of deferred access by a document delivery system.

The model of documents in the data warehouse is based on the separate distributed storage of the document text and its descriptive metadata, based on international standards and recommendations. This type of storage allows distributed retrieval and storage of electronic documents at the point of production.

Although all document access technologies share common origins, in the process of their development, due to their different fields of use, they have been divided into three independent directions, determined by the nature and location of document storage.

To meet the information needs of modern society, which is characterized by the use of a large number of approaches to integration with existing service systems and the definition of functional tasks, it is necessary to analyze approaches to the creation of departments that serve users with full-text electronic documents.

Another problem of the organization of effective information support for scientific and educational activities is that, due to its diversity and multidimensional nature, scientific and educational information resources are dispersed on remote pages of many sites and in distributed digital libraries and archives.

Distributed information systems to support scientific and educational activities operate with various kinds of information. These can be publications, electronic documents, electronic collections, ontological descriptions, data arrays, logical descriptions, etc. These resources, demanded by different groups of researchers, may not be available due to problems with their search and identification. Semantic links between information resources increase their value and provide additional opportunities for information retrieval and identification.

The support of distributed information systems in the field of scientific and educational activities is relevant, since the need for information always exists. To satisfy this need, it is necessary to organize access to various resources.

2. Literature review and problem statement

The paper [1] describes the role and features of the use of automated information systems in the activities of universities, and also gives a general description of existing information systems in the field of education. The concept and approaches to the introduction of electronic educational information services for students, teachers and administrative staff, as well as an information system for university management in the Republic of Uzbekistan are described. The information management system of the university includes all tasks related to the administrative, scientific, scientific, financial and economic activities of educational institutions. The paper examines the specifics of managerial and educational processes in Uzbekistan. As a consequence, business process and workflow reengineering has been developed tak-

ing into account administrative and business management aspects. Finally, current research shows that traditional education management processes with promising information systems capabilities can be transferred to electronic services.

The paper [2] describes the developing technology in the field of databases that makes it possible to directly manage semi-structured documents in accordance with the requirements of business processes. In a business environment, there may be a number of possible document formats. Documents can be classified according to the organization of the basic data sets and according to the need and capabilities of the data included in the documents, depending on whether they should be fully or partially structured. The most advanced database technology provides tools for processing and extracting data using semi-structured and unstructured data.

The work [3] is devoted to the creation of a conceptual model of an information system to support scientific and educational activities. The paper discusses the information needs of the modern user and information objects that describe the main entities of the scientific information space, such as publication, document, person, dictionary entry, function, and user, as well as the connections between them.

The paper [4] describes access to system information resources of scientific papers. Therefore, in any field of science, it begins with research, the search for scientific information, but with the increase in the number of scientific papers, books, monographs, patents, the search for information becomes more and more difficult. Creation of a unified information system allows scientists to quickly get acquainted with the results of other scientific research and prevent their duplication. The paper discusses the technological techniques of distributed information systems that provide scientific and educational activities. The main tasks of creating a model of a distributed information system supporting scientific and educational activities, the functionality of the model, the concept of metadata and the requirements for the metadata profile are described. The task, subject area, subjects, objects, the main functionality of the information system are defined, a list of the main types of information resources is given. The work analyzes the functional requirements for such systems.

The paper [5] describes the potential of the scientific libraries of the city in the information support of the objects of the scientific and educational complex and shows how the transformation of the scientific and educational space affects certain areas of activity of scientific libraries. The existing ideas about the component structure of interdependent subjects of the educational sphere are updated. The potential of scientific libraries is determined, their documentary resources are characterized, taking into account social factors affecting the structure of the scientific and educational environment of Novosibirsk. It is argued that the accelerated integration of large universities and research institutes encourages scientific libraries of different status to develop new models of corporate interaction. With the disproportionate provision of scientific libraries of the city with information resources, it is advisable to form an adequate model of interaction between all participants in the scientific and educational process and create, based on the use of navigation tools, a system for informing about the resource capabilities of the entire conglomeration of scientific libraries of the city. The convergence of the resource base of the scientific libraries of the city may have a compensatory value for the information support of scientific and educational activities.

In [6], a comprehensive approach to building a model of competencies of a specialist in library and information activ-

ities is proposed, based on a combination of approved regulatory documents and employers' requests: for each group of competencies, a basic minimum (approved by federal bodies responsible for state supervision, and ensuring compliance with the requirements for the composition of labor functions, knowledge and skills in approved documents) and a variable part should be determined, ensuring compliance of the formed competencies with the peculiarities of the activities of libraries of various types (public, academic, special).

This requires the development of a regulatory framework in the field of library and information activities (updating existing and developing new documents). The proposed competence model is correlated with regulatory documents. A competence model is proposed, within which educational institutions, libraries and information institutions can jointly develop an effective process of training specialists.

The paper [7] examines the modern infrastructure of libraries and information. The development trends are shown, the place of scientific libraries is determined. The main directions of library and information activities aimed at improving the quality of information support for the scientific, educational, scientific and industrial sectors are highlighted. The main national sources of information are described. The most important areas that require rethinking and inclusion in the special work of academic libraries are: the study of document flow in order to optimize the processes of forming collections with the participation of bibliometric and webometric services; rational localization of information resources in research centers and access to them online; work with free and open resources. For scientific libraries, the use of bibliometrics is important, as it allows you to competently build a strategy of acquisition and thematic information on models of current and selective dissemination of information based on citation analysis by statistical processing of the frequency distribution of references by publications and authors.

In [8], based on the analysis of typical scenarios of information servers, the tasks that should be solved when organizing an access control system for distributed information resources are formulated. The possibilities of the Z39.50 technologies as the most suitable for building such a system are considered. Within the framework of this technology, three access control models are discussed, which differ in the degree of integration of information server functions with the Z39.50 technologies.

The creation and support of distributed information systems and electronic libraries that integrate heterogeneous information resources and operate in various software and hardware environments require special approaches to managing these systems. If the resources or data themselves can be managed locally, even for distributed information systems, then the task of managing access to distributed resources cannot be solved within the framework of local administration. The justification of the last thesis can be seen when considering typical scenarios of the information server, which we will describe below.

After studying the above-mentioned works, we decided to create models of documents that support scientific and educational activities.

3. The aim and objectives of the study

The aim of the work is to develop technological solutions specific to a research organization for providing access to documents that allow both unification of access to heterogeneous information resources and integration with other information systems.

To achieve the aim, the following objectives were set:
– to develop a model of information processes of document access technology;
– to define a data storage model and descriptive metadata schema of library documents.

4. Materials and methods

Large volumes of already existing information, its continuous growth, the heterogeneous nature of storage and distribution, the lack of a unified way of working with it create many difficulties when using it. Awareness of the mentioned difficulties, qualitative changes in the field of information technology and telecommunications have led to the need to solve the problem of finding new approaches to creating repositories of information resources, their structure, and the development of tools necessary for users.

To meet the information needs of modern society, which is characterized by the use of a large number of approaches to integration with existing service systems and the definition of functional tasks, it is necessary to analyze approaches to the creation of departments that serve users with full-text electronic documents.

Depending on the nature of access to documents and the source of their occurrence, the following groups of documents and their corresponding information environments are distinguished:
– Documents stored in libraries. They can be in electronic or another form, depending on the material medium (hard copy, microfilm, etc.).
– Documents issued by publishing centers (publishing houses, subscription agencies, etc.). There are both electronic and printed.
– Documents hosted on computer networks. They exist only in electronic form.

The user often needs access to documents related to various groups, which has led to the development of information resources and systems for working with them, combining the achievements of each of the development directions.

The presentation of information and knowledge not only in traditional printed but also in electronic form is a distinctive feature of the present stage of development of society. This allows you to create, store, use the information and organize access to it in a fundamentally different way. Modern information technologies make it possible to transfer accumulated information into electronic form, as well as to create new information resources directly in electronic form. As a result, there is a construction of a space in which a huge amount of heterogeneous information functions. The totality of electronic documents and information systems that ensure the functioning of these documents is an electronic document space, there is a rapid increase in collections containing electronic documents. This led to the need to rethink the role and functions of libraries in an electronic environment. The high rates of information accumulation have made it necessary to search for tools that allow for quick and effective access to knowledge dispersed across various information repositories. One of these methods is the technology of electronic libraries. An electronic library is an ordered collection of electronic documents intended for long-term

storage and public use, formed under certain criteria [9]. Search through the entire text of an electronic document expands the possibilities of finding relevant sources in huge amounts of information. Currently, an electronic document is recognized as an object of library activity. In this regard, specialized collections are being created in the structure of library collections, allocated according to the form of presentation. They are called electronic collections, electronic libraries, and electronic funds.

5. Results of research on the development of a model and technology of access to documents in scientific and educational activities

5. 1. Development of a model of information processes of document access technology

We consider access technology as a tool capable of improving the quality of information by ensuring its safety, creating accessibility conditions, primarily search and delivery, in a user-friendly form and a convenient place for him. Consequently, the functions of the developers of the technology of access to documents of a specific subject area include, in addition to the actual access, ensuring the selection of material and the creation of thematic structuring of the collection of information resources.

The listed definitions of access indicate the differences between access to documents in an electronic library and Internet search engines. The user of the access system deals with carefully selected high-quality information resources, the relevance of the search is provided by the keyword system, the contents of the repository are presented in the form of a well-organized database structure, which is annotated and constantly monitored by specialists. As a result, the user receives a relatively small compact array of information, which has three levels of detail:

a) a tree or a list of information resources;

b) metadata about documents;

c) the referenced document.

Thus, the technology of access to documents can be defined as a complex system that includes three mandatory components:

1. System users.

2. Document access block.

3. Document storage.

Such a representation of the system determines the need to take into account all three components when formulating requirements for access technology.

It should be noted that a research organization has many features that are essential when developing technology for accessing documents intended for its employees [8]:

1. An extensive range of information sources is necessary for research activities. A poorly organized array of such sources leads to significant time spent searching for them.

Users need a single interface that integrates all the resources available to the organization in a single window.

2. Limited thematic coverage of documents determined by the field of research of the organization. This feature implies the need to allocate such resources from the entire set of information offered by suppliers.

3. Highly qualified users. This feature assumes that users are aware of the latest methods of accessing documents and desire to use them in their activities.

4. The need for mutual exchange of scientific information causes interest in open archive technologies. In addition, users need the means to publish their copyrighted works on the Internet to inform third-party users about the progress and results of research.

5. Administrative relations of a research organization as a structural subdivision of a regional scientific center, a branch of the Academy of Sciences, and the entire academy as a whole. The presence of such links implies mandatory interoperability and good integration of software and technological solutions into the general environment of the Academy of Sciences.

6. The minimum staff of developers implementing access technologies and their insufficient funding due to the secondary nature of information support services for scientific research in the organization. This feature determines the need for widespread use of ready-made software solutions and the ease of their maintenance.

7. The presence in the organization of a large number of electronic documents of its products that are not subject to licensing, as well as documents obtained within the framework of various scientific and charitable programs, as well as personal connections. Collections of such documents from employees of the organization form a fund of so-called personal collections of documents requiring inventory.

Access to them is desirable because of their scientific value and can be organized based on the goodwill of their owners. This procedure requires careful differentiation of access to documents not only at the level of collections of documents but also each document individually.

After a preliminary analysis of the development of access technologies, highlighting the necessary components and taking into account the specifics of the research organization, the following basic requirements for access technology can be identified:

1. Openness and extensibility. Access to documents from the Internet is based on open system standards, which ensures the portability of technological blocks to other hardware platforms, as well as the connection of information sources from other manufacturers. It is provided by the use of Web technologies, RUSMARC communication format, and Z39.50.

2. Network protocols. Ease of management. It is achieved by using functional modules in the technology, determined by the commonality of information processes and the information technologies used. Splitting into modules allows you to use ready-made software solutions from different developers. Modularity allows for autonomous maintenance of each unit, as well as upgrading and replacing the software used in them as the relevant information technologies develop.

3. Centralized user interface. The document is searched and accessed from a "single window". For the user, the system looks centralized, since its distribution is hidden by the interface.

4. Distribution of documents and collections by storage locations and decentralization of their administration. Documents, collections, and groups of collections are stored where they are formed.

5. Database-level integrability. The technology provides access to resources from various manufacturers and also allows logical grouping of collections of documents, including those created using various DBMS, to form thematic resources and resources arranged by region.

6. Differentiation of access rights. To ensure the terms of license agreements and copyrights, the level of access to the organization's databases and individual documents is determined by the individual rights of a user or a group of users.

7. Unified descriptive metadata and provision of multifunctional search on them. To provide access to documents

using various search interfaces, documents of various types are presented in the system in a unified way of describing two types that allow mutual conversion:

– in the metadata generation block, the document is described by the ABC in the MARC family format;

– a document entering the document repository through the expert repository receives an RDF description in the Dublin Core (DC) or MODS schema.

8. Unified document storage formats. The document storage format should allow its semantic and syntactic analysis for subsequent search in the full text.

9. Variability of formats for displaying document search results on the screen – from MARC formats to the format of a catalog card. Access to an online electronic document is encapsulated in the output format as a URL link.

10. The possibility of "deferred access", i.e. registration of the application, conversion of the document into electronic form with preservation, and subsequent delivery to the user through the access system or by email.

Taking into account the formulated access requirements in the three-component scheme of document access technology as a system, it is possible to identify modules and determine technological information flows. The information that the technology works with is divided into permanent storage information, such as documents and their metadata grouped into arrays, and temporary storage or dynamically generated information – user requests, orders, and messages about alternative ways to receive it, about the unavailability of the document or consider the general logical scheme of the technology of access to documents (Fig. 1). Input information is presented in the following types: information about the set of databases required by the user for search, user requests, full texts of documents to be stored.

Output information – full texts of electronic documents or information about their absence, as well as alternative methods of obtaining, if the document is printed. There is also an additional information flow to replenish the document repository through an intermediate expert repository, bypassing the document access block. The documents contained in the repository are represented in the document access block by descriptive metadata contained in the metadata repository. The user forms a set of databases and a search request through a single resource access window, the request is sent to a search engine that searches the metadata repository, returning its results to the access point. If the document is detected, the user requests it to the repository using the delivery module and receives the electronic document to the access point. The metadata storage is replenished by describing the document in the metadata generation module.

Then the document description is assigned to some metadata array (or duplicated into several arrays) and enters the metadata repository.

The technology also provides linguistic search support by forming search dictionaries based on document metadata elements.

In the logical scheme of the technology, functional modules that implement certain information processes are distinguished, and modules of information resources of the technology – data storage.

A functional module is a technological unit of a target purpose with its information flows. The allocation of functional modules is due to the following signs of their internal unity:

1. The uniformity of the information technologies used within the module and the means of their implementation.
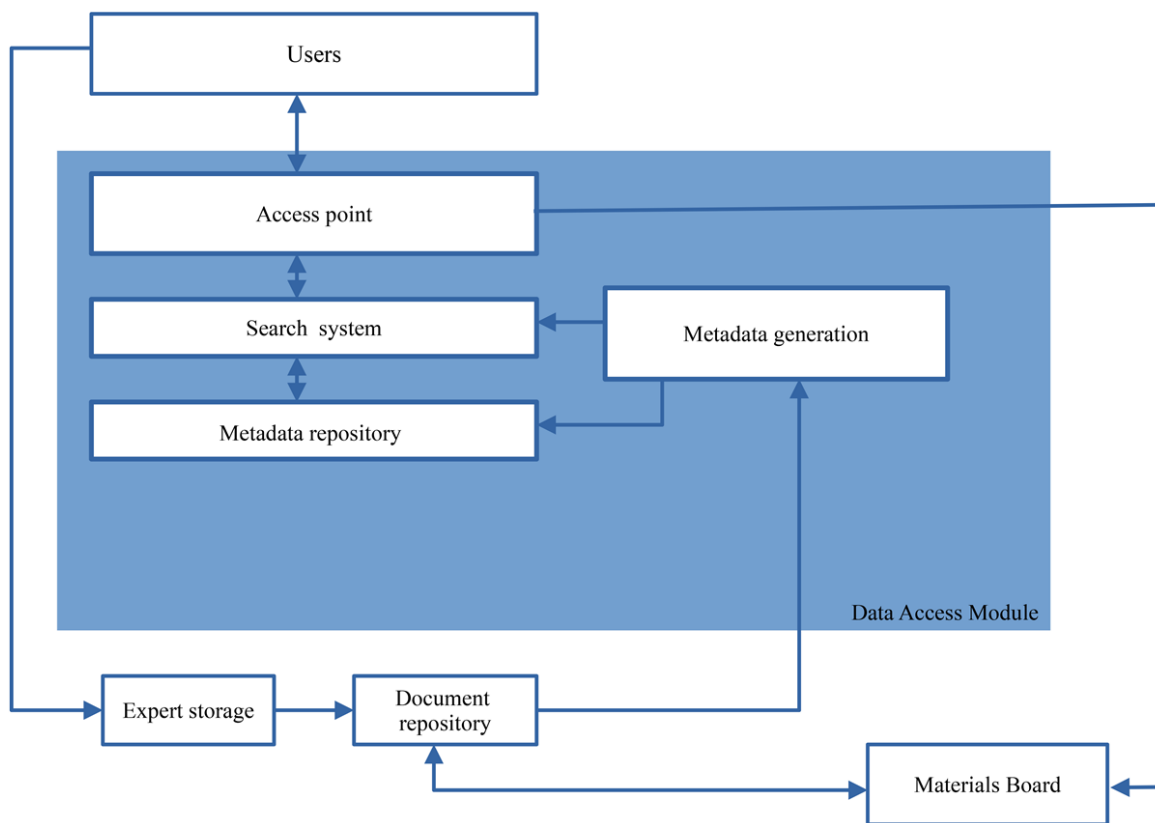

Fig. 1. Logical scheme of document access technology

2. The need to attract specialists for the modernization and maintenance of each unit.

Permanent storage information resources are divided into two types:

1. Documents.

2. Descriptive metadata of documents.

Each type of resource is allocated its storage following the three-level model of technology as a complex system [8].

Let's take a closer look at the functional modules of the technology, define information flows and functionality requirements.

A single access point. The resource access point implements the "single window" technology and is intended to serve as a user interface for accessing documents. From the definition of requirements, it follows that the access point will be used to work with Internet resources

Fig. 2. Logic diagram of the access point module

published under the management of Z39.50. The segment of the Internet managed by Z39.50 is much more organized due to the traditional use of protocols by the library community and therefore preferred when searching in arrays of descriptive metadata of documents. Access to information under the control of Z39.50 is carried out in two ways: using the z-client and using the Z39.50-HTTP gateway. The use of the client undoubtedly has many advantages, first of all, allowing you to organize access to metadata without any intermediary data transfers between the environments of the two protocols. In addition, the existing gateway implementations do not provide full-featured use of Z39.50. However, the use of the z-client requires the installation of client applications on each of the user's computers. Given the rapidly growing number of computers used by employees of a research organization, it is very costly. In addition, the transition from metadata to the URL link of documents stored in the Web environment in the original creation formats requires separate efforts. The gateway allows you to organize access to data in a Web environment familiar to users using standard browsers that operating systems are equipped with. The use of a gateway with a subsequent refinement of its functionality in this situation is more preferable [8].

The development and implementation of a document access point are carried out based on Web technologies, representing a Web interface designed to provide access to documents. The access point is designed in the form of a Web page containing a thematically organized list of available resources. Such an organization is intended for the user to select the resources for which he needs to conduct a search, to obtain information about these resources, as well as the wording of the search query and transfer it to the search engine. The results of the search query are also transmitted to the access point to display them on the display screen. With the help of an access point, the user is informed about the databases available to him, and he can independently allocate a list of resources to search for the document he needs. The logical scheme of information flows is shown in Fig. 2.
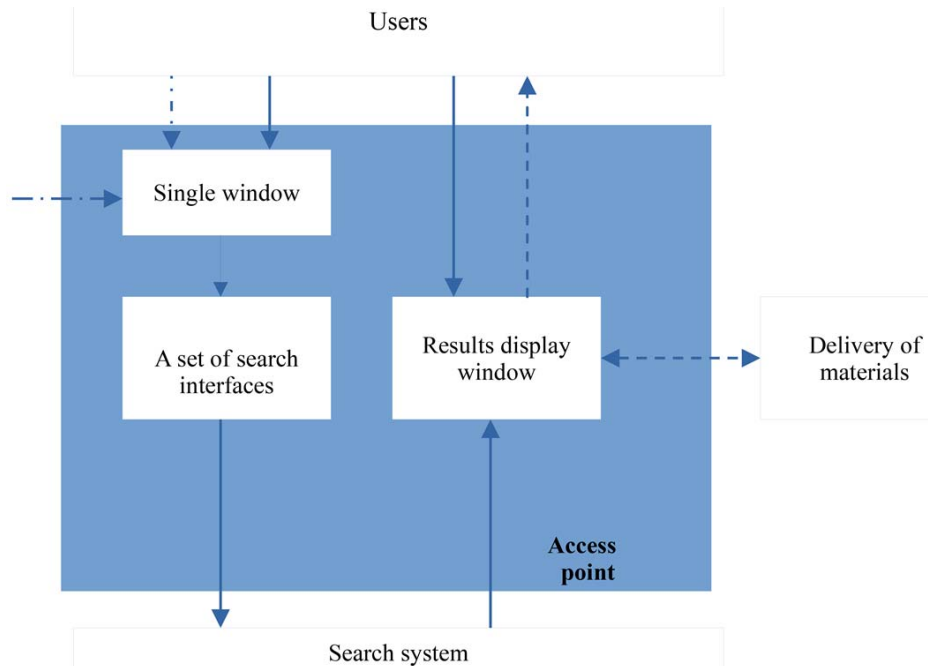
The access point operates with three information flows:

1) user search queries with the vector "user – single window – set of search interfaces – search engine – results display window – user";

2) information about arrays of meta descriptions with the vector "user – single window – search engine – single window – user";

3) requests for document delivery with the vector "user – results display window – delivery – results display window – user".

It is clear from the diagram that the input information for the access point is a user request, the contents of which the access point transmits to the search engine through a set of search interfaces. Also, the incoming flow of information for the access point is the result of the request, forwarded to the user using display formats or a message that the search result is empty. In case of a positive search result, the user also receives information in the output form about the presence or absence of an electronic version of the document, which he requests through the delivery module. The request to the delivery module is the output information stream of the access point. An electronic document is also returned to the access point here to display it on the display screen.

Search engine. The basic requirement for a search engine is the possibility of distributed search through metadata arrays posted on the Internet. Descriptive metadata of documents are created on different platforms using different DBMS. To integrate such data into a single environment, the Z39.50 network protocols (ISO 23950 standard) are used. The Z39.50 standard is one of the protocols of the OSI family, which describes the application level of interaction of distributed information retrieval systems. The protocol defines the mechanism of information exchange in the process of processing search queries and the protocol of data exchange in the systems performing the search. The main area of application of the protocol at present is library systems and scientific and technical information systems. However, the scope of the protocol is much wider than the listed applications – it can be used in general-purpose in-

formation search engines. When developing the protocol, it was assumed that it would describe the procedure for exchanging information between users of an information system and its core through a data transmission network. At the same time, the systems themselves can manage data using different models and different languages for manipulating this data, whether it is a regular file system or an object-oriented DBMS [9]. The Z39.50 protocol does not define the transport layer of interaction in the network and can be implemented on top of any transport protocols, for example, TCP/IP, used for communication on the Internet. The protocol describes network interaction in the "client-server" architecture, using the concepts of "origin-target" to determine the relevant subjects of interaction. The transfer protocol focuses on a persistent connection between origin and target, called a session. When a session is opened, special session variables are dynamically created in target, which is destroyed when it is closed. Session variables store information related to the current session: query history, settings, user information, etc. They store named datasets that are available for use in repeated requests [9]. This feature of the protocol allows you to meet many other requirements for DIS:

– saving and reusing queries;
– the possibility of a clarifying search;
– sorting search results;
– presentation of results in various formats with the ability to save records.

In Z39.50, search queries are always formulated not to a real database, but an abstract one. This abstract database, called a "set of attributes", has no structure and is characterized only by search attributes.

With this approach to the search procedure, all databases become the same for the user if they support the same set of search attributes. For a subject area with similar search attributes that stand out from a standardized description of objects, for example, the MARC format, such a search model seems to be a very good choice. The data extraction model in Z39.50 includes the mapping of result set records to abstract database records through a schema defining the abstract structure of the record. Sets of search attributes and data schemas are standardized for different protocol applications.

To work with bibliographic meta-information, a special set of bib-1 search attributes and an ISO-2709 data schema corresponding to the MARC family of standards have been defined. But the use of Z39.50 is not limited only to bibliographic data but defines schemes for describing information on the Web (Dublin Core, DC) or describing digital collections. This once again confirms the prospects of using the protocol in document access technology as objects of the Web environment. The use of Z39.50 in the search engine module determines the compliance of the technology with the requirements of openness, extensibility, interoperability, and distribution.

The logic diagram of the module (Fig. 3) contains two functional blocks and an IR-Explain database of descriptions of metadata arrays. Three information flows have been identified:

1) user search queries with the vector "access point – search block – access control – metadata storage – search block – access point", or in case of access denial "access point – search block – access control – search block – access point";

2) information about metadata arrays with the vector "access point – IR-Explain database – access point";

3) linguistic support of search with the vector "metadata generation – search block".
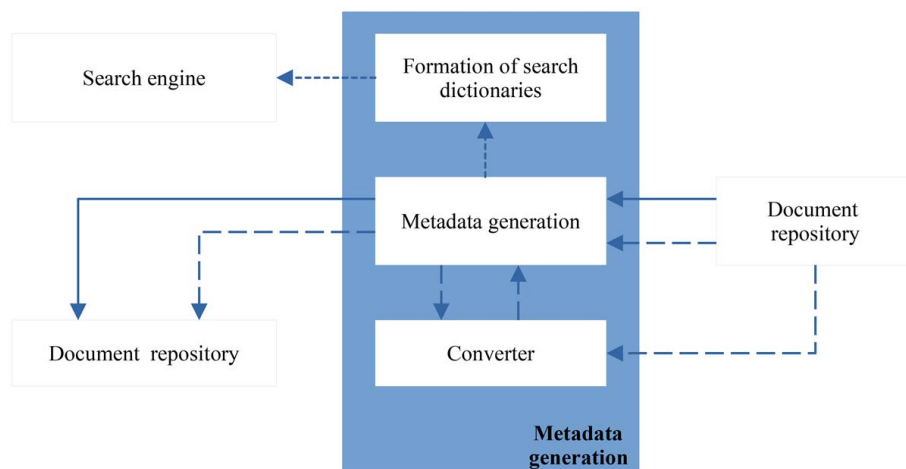

Fig. 3. Logic diagram of the search engine module

The IR-Explain database, a special tool Z39.50 containing a meta description of data arrays, is an integral part of the protocols, uses Z39.50 tools to extract data, and therefore is also referred to as the search module of the system based on the commonality of the software used. The module's access control unit checks access rights to restricted metadata arrays by the value of the user's IP address.

Metadata generation. The metadata generation module is designed to process the full texts of documents before placing them in the repository. During the processing of the document by the metadata generation unit, a secondary information resource is created, called a bibliographic description of the document. The presentation of the bibliographic description, as well as the results of indexing and abstracting the document in the standard machine-readable form of the MARC family, creates a bibliographic record or a search image of the document that can be used by the search engine. Bibliographic records are compiled into named arrays of records, indexed, converted into documentary databases, and placed in the metadata repository. Any automated library and information system used in the organization's library is suitable for this, under one condition – a Z39.50 data provider must be written for it, displaying the EduDIS data in an abstract data schema.

However, access to a set of documents in the repository cannot be limited to a single type of access point. To provide access to documents to the widest range of the scientific community, a Web search engine or E-print open archive technology can also be used as an access point to the document repository. For example, to present a document in an open archive, it is necessary to describe it in the DC standard using the RDF specification, which can be stored either in the document body or separately if the document is not in hypertext format.

As a consequence, the metadata generation module performs two more functions:

1. Providing the search system with dictionaries formed from the semantics of the document using a set of various

predefined indexing methods applied to the fields of descriptive metadata of the document. This is also a function of the ABIS used.

2. The metadata generation module should have a set of converters for translating metadata from standard to standard.

Converters can be embedded in the EduDIS, or they can be a separate software product that avoids additional work on duplicating metadata into different standards.

The module contains three functional blocks and serves the following information flows (Fig. 4):

1. Formation of descriptive metadata in MARC format with the vector "document storage – metadata generation – metadata storage".

2. Metadata generation in the Dublin Core format with the vector "document storage – metadata generation – converter – metadata storage". A counterflow is also possible for converting metadata in DC format contained in an electronic document into MARC format with their subsequent correction, with the vector "document storage – converter – metadata generation – metadata storage".

3. Linguistic support of search with the vector "metadata generation – formation of search dictionaries – search engine".

Delivery. The operation of the delivery module depends on the type of required document and access rights to it:

1. If the electronic version of the document is publicly available, then the delivery module simply provides it to the user's computer via a URL link.

2. If access to the electronic version is restricted, the user can place an order for delivery and, if possible, receive it by email.

3. If the document exists in printed form and its conversion to electronic form is possible, a delivery order is registered, and the resulting document is sent to the user by email. In case of a great need for this kind of service, it can be automated using one of the automated electronic document delivery systems, for example, [10].

The logic diagram of the module (Fig. 5) contains the functional blocks of the module and defines the following information flows:

1. Delivery of an electronic document with the vector "access point – access control – document storage – access point".

2. Delivery of an electronic document with access restrictions with the vector "access point – access control – access point – order formation – user".

3. Delivery of a printed document suitable for digitization with the vector "access point – order formation – production of an electronic document – user".
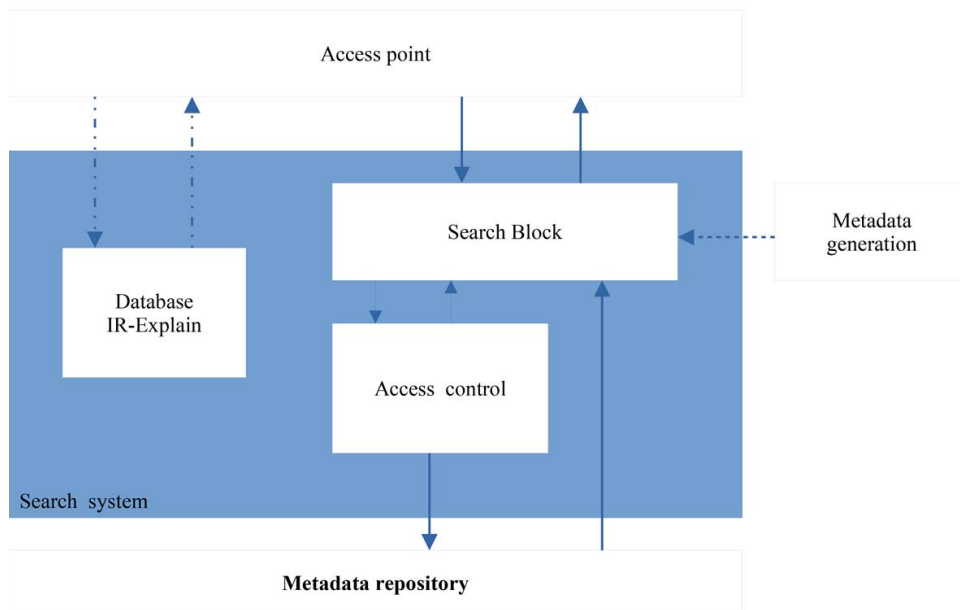
Fig. 4. Logical diagram of the metadata generation module
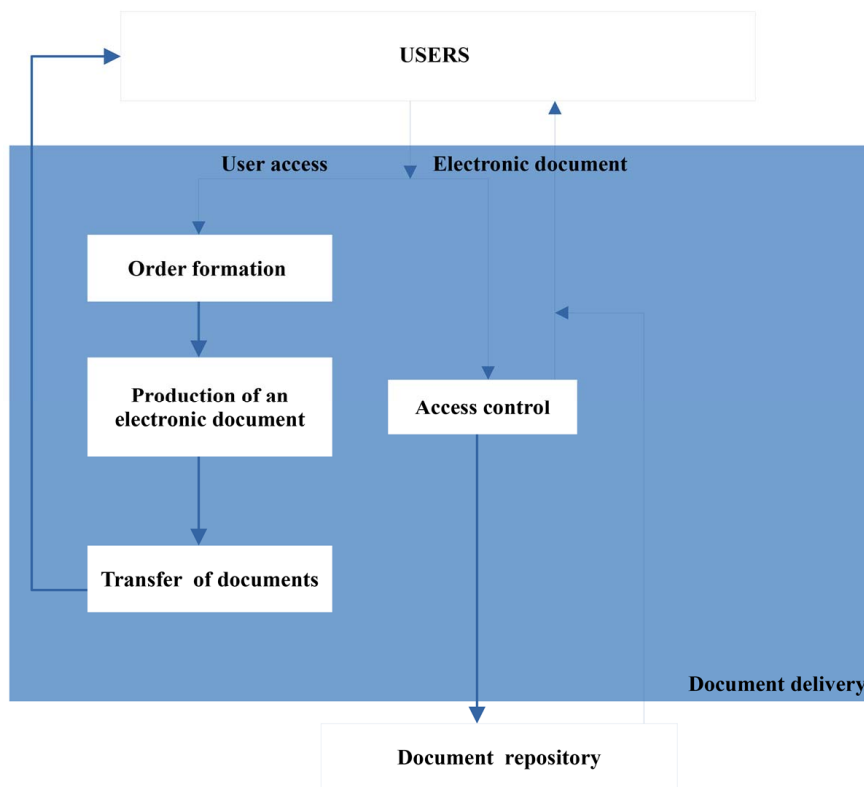
Fig. 5. Logic diagram of the delivery module

Within this module, the ability to deliver an electronic document is determined by the access rights to it that are set during the production of the document at the file level or the

server directory level employing a Web server. Web tools are used to deliver the detected document. This determines the need to store documents in a format suitable for their display by the browser window.

Thus, the data model of information processes of document access technology can be represented as an interconnected complex of functional modules (Fig. 6). The division into modules is determined by the functionality and the software used in each of them.

The use of a distributed repository of documents and metadata, as well as an object model of the document, makes it possible to form the description of a wide set of documents. The versatility of the data model used allows you to vary the software modules, creating the desired combinations.
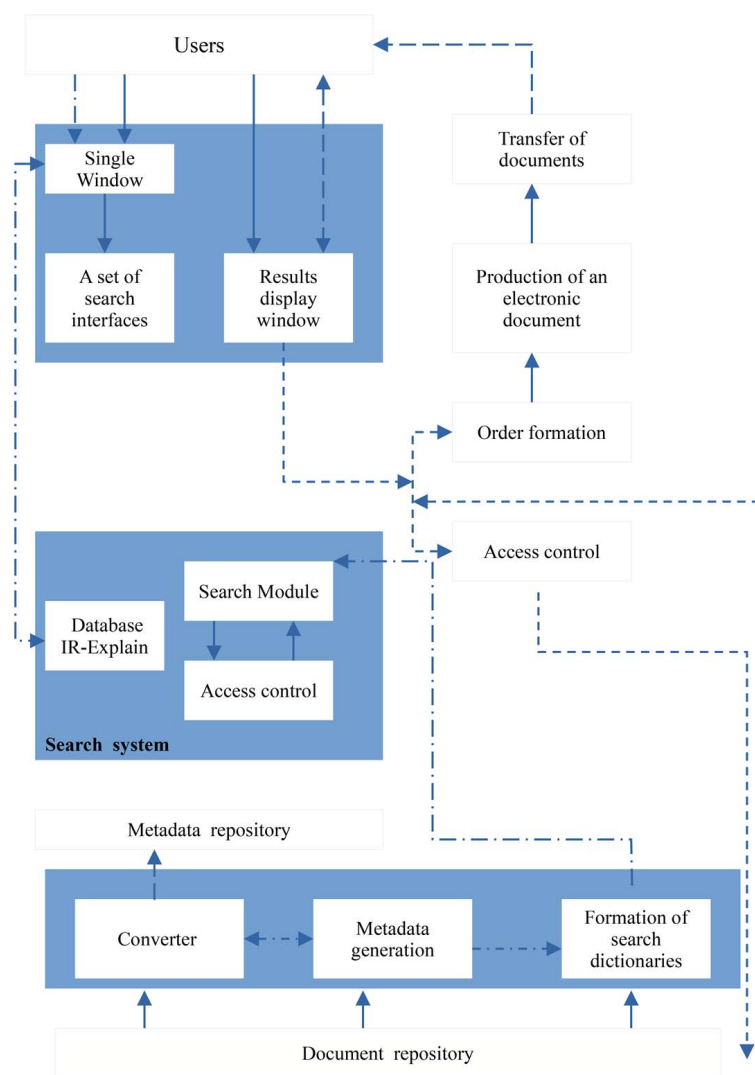
Fig. 6. Model of information processes of access technology

If the document exists only in printed form and cannot be converted, then there is simply no means of calling the delivery module at the access point. But in the results window, there is information about the presence or absence of a document at the place of its storage. To provide this function, a complete electronic catalog of printed documents of the repository is required, and the function of electronic book issuance with the registration of the fact of issuance in metadata is implemented.

5. 2. Data storage model and descriptive metadata schemes of library documents

For a more convenient presentation of information to the user, the visualization of knowledge and data stored in the EduDiS content is configured. When configuring visualization in the ontology editor, a template for visualizing objects of this class and a template for visualizing links to them are set for each class.

The visualization template for class objects (information objects) includes all the attributes of this class and the relationships associated with it. There are "direct" (directed from this class to other classes) and "reverse" (directed from other classes to this) relationships. When visualizing classes and information objects, relationships are grouped by these two types.

By default, class attributes and related relationships, including relationship attributes, are displayed in the order in which they are specified in the ontology. At the user's request, this order can be changed.

The required order of attributes is set using a specially allocated property (annotation property) called to order. The value of this property is a number that specifies the position number in the attribute sequence. These properties are set for simple values (datatype property).

A template for visualizing a reference to an object of a class can include both the attributes of this class and the attributes of the classes associated with it and the relationships defined between it and other classes. There are two types of links – full and short. Full references are used when displaying a list of instances of a given class, short references are used when referring to an instance from another instance. For full links, the link property is highlighted, for short links – a short link. The value of these properties is also a number that specifies the order of the components in the link. These properties can be set both for simple values (datatype property) and for object properties (object property), i.e. relations.

The values of the attributes included in the link are used to build a text representation of the link to an object of this class when it is displayed (visualized) on the screen.

The EduDIS creation technology provides the ability to customize the display of information objects included in the EduDIS content (instances of ontology classes) when they are displayed on the monitor screen (Fig. 7). For these purposes, it offers a mechanism for presentation patterns [11] that allows you to customize the visualization of objects of selected classes when they are shown to users and edited. To implement such patterns, special order, link, and short link properties were introduced into the ontology, which serves to set the order of displaying attributes during visualization and forming short (shortlink) and full (link) references to objects. (A reference to an object is its name, made up of the values of its properties and the properties of the objects associated with it). These properties are declared in the ontology as Annotation Properties, which allows you to store information about content visualization rules directly in the ontology.

ИНФОРМАЦИОННАЯ СИСТЕМА ДЛЯ НАУЧНО-ОБРАЗОВАТЕЛЬНОЙ ДЕЯТЕЛЬНОСТИ

Fig. 7. User interface

To fill in EduDIS content, information is collected from sources such as websites of organizations, associations, projects and conferences, knowledge portals, social scientific networks, etc. Information about Projects, Organizations, Personalities, Conferences is extracted from these sources, i. e., about all the basic classes of the ontology of scientific activity, except for information about Publications. Information about Publications is extracted from the repository (Dspace), which was created by the authors.

For each of these classes, a different method of information extraction was created, including a set of templates generated based on the ontology. To increase the completeness of information extraction, the variability of these templates is increased by using alternative terms from the thesaurus (synonyms and hyponyms), as well as words and phrases proposed by the expert.

Documents on the Internet can be presented in various formats (HTML, DOC, PDF, TXT, and others). The main format for presenting information on the Internet is HTML. To extract publication metadata from repositories in batch mode, data is exported in XML format (Fig. 8). The proposed methods of extracting information about Projects, Organizations, Personalities, Conferences are focused on working with HTML pages, and information about Publications is focused on working with XML documents.

To facilitate the analysis of HTML pages and XML documents, the resource is represented as a DOM tree under the DOM (Document Object Model) standard, which regulates the way the contents of the document (in particular, HTML pages and XML documents) are represented by a set of objects [11]. Based on the corresponding template, the DOM tree of each page is analyzed and the information described by this template is extracted.

It is important to note that information about entities of interest to EduDIS users can be set in various ways. For example, information about the project can be provided by the project site, a section of the organization or person's website, or a publication describing the project. For each of these ways of representation, a separate template is built based on the ontology class of the Project.

Document model in storage.

A document used in access technology is understood to mean a primary document in printed or electronic form. A document in electronic form exists as a file in a certain format, which requires special software tools to view it on a computer. Let's consider the document d_i-$<p_i, m_i>$, where p_i is the content part (or file) of the document, m_i is the metadata of the document. Equivalent sets $P=\{p_1, p_2..., p_n\}$ and defining the array of documents $D=M=(m_1, m_2, ..., m_n)\{d_1, d_2...d_n\}$ form two related data stores: a document (or full text) repository and a metadata repository.

A document repository is a distributed repository of the full texts of documents, not necessarily electronic. The full text of the document can be a printed document located in the premises of the library of the organization. Electronic documents can be stored both together with their metadata and separately from them, including on a remote server.

The impossibility of integrating heterogeneous documents defined in this way in an electronic environment determines the use of secondary documents – bibliographic records (or descriptive metadata) to search for the document and initially inform users about its semantic content. Thus, document metadata, from the point of view of the Web environment, are documents and the document object model is fully applicable to them.

These definitions refer to the general case of modeling documents and the relationships between them when

working with information resources that belong to different classes of documents. Let's define the same concepts in relation to a single class.

Definition 1. The metadata of the document m_i will be called the pair $m_i = \langle S_i, V_i \rangle$, where S_i is the metadata structure in accordance with the selected schema; V_i – content of the scheme.

It follows from the definition that metadata is divided into descriptive and structural. Structural metadata (structure) defines the structure and properties of documents according to which they are processed (types, relationships, presentation formats, access control restrictions, etc.). Descriptive metadata (content) describes the semantic content of the document (its title, summary, etc.). Depending on the data schema you choose, descriptive metadata can contain information about the structure and properties of the document (as in MARC format).

Descriptive metadata can be part of an electronic document or stored separately from it.

The element of the data schema will be called its structural element.

Metadata generated in different data schemas will have a different set of structural elements. Any structural element of the data schema that defines the edition typically consists of an identifier and the name of the element. For example, a MARC schema uses a digital identifier and the name of a field and subfield. And in the Dublin Core scheme, there are 15 basic elements, defined only by their names. To determine the identity of two documents from their bibliographic record, a comparison of the contents of the set of structural elements selected from the diagram, different for different types of documents, is used. Various comparison methods are often used for this purpose, for example, comparison by signature or the method of fuzzy string matching.

Define a minimum set of data schema elements that uniquely identify a document as a document comparison index. Let's consider the metadata that matches the content of the comparison index to be identical.

Definition 2. Two documents d_1 and d_2 will be called document instances if they have identical metadata $m_1 = m_2$.

Definition 3. Collection is a set of documents (or their metadata) with a dedicated fixed structure, the content of which has the same thematic focus.

It is clear from the document definition that for the set of documents D in the document store, the concept of a collection is blurred, since the definition of the document does not explicitly include a description of its structure. The concept of a collection is defined on a set of metadata. Due to the connectedness and equivalence of the sets D and M, each element of the collection m_i, the set M, corresponds to the element d_i, the set D. In the MARC format, for example, the relationship of the description to the document is determined by the field of reference to the document.

Thus, the entire set of metadata in the metadata repository is divided into collections. In other words, the collection Kj is a subset of the set $M : K_j \in M$. In general, the entire set M is a collection of publications available to employees of a scientific and technical organization.

All the set of collections is distributed by places of their formation in the network.

Collections can be logically combined by some feature, for example, by the type or nature of documents, the organization of formation or the geographical location of the collection. In general, any of the K_j collections can be a distributed information resource. The set family K_j is a partition of the set M, where J is the set of indexes and $M = U_{j \in J} K_j$ do not overlap in pairs. It follows from the statement that there is no metadata in the store that does not belong to any collection.
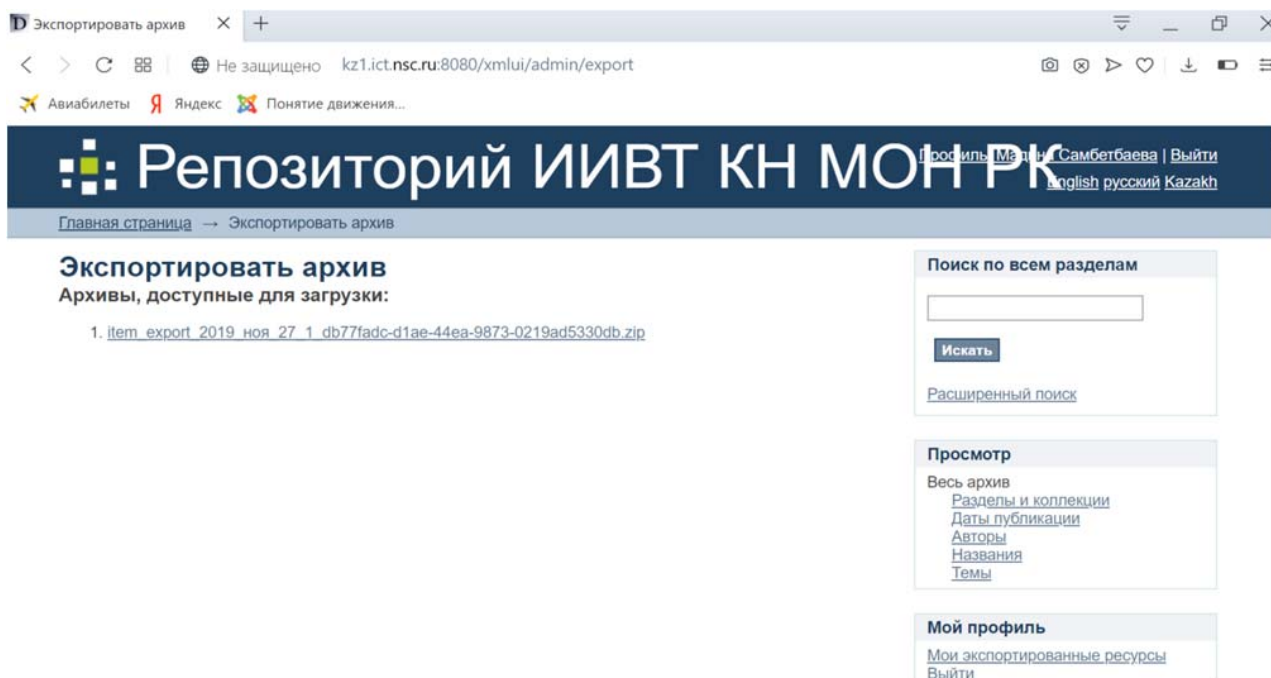
Fig. 8. Exporting metadata from the repository

Within a single collection, each document exists in a single copy. However, there can be descriptions of instances of the same document in the metadata store that belong to different collections.

Copies of a document can exist due to the multiplicity of both printed documents in the collections of libraries and the multiplicity of copies of electronic documents in different storage locations. Metadata of document instances is uniquely defined only by different places of their storage; the content of all other structural elements can coincide. If a group of collections of publications has a similar thematic focus, or the organizations that form collections of publications of employees have close scientific ties, then in these collections there may be a sufficiently large number of matching metadata that define the copies of the document. When the number of such matches is comparable in order of magnitude to the size of the collection, this indicates the desirability of merging such collections, accompanied by merging the metadata of the document instances. In this procedure, you treat the metadata of document instances as the intersection of many different collections in which they are contained. When you merge sets, the defining collections merge in pairs until all the combined sets are exhausted. In this case, the symmetric differences of two sets of source collections are copied into the resulting set, and the set of matching copies of documents is treated as the intersection of the sets [12].

This compares the contents of all structural elements of document instances, copies the contents of all structural elements to the resulting element if a match, and another instance of the structural element is created in case of a mismatch.

The resulting collection forms a consolidated information resource. When you create aggregated metadata collections, multiple documents and their metadata in the repositories are no longer equivalent because the metadata store generates items that are associated with document store items in a one-to-many relationship. Metadata appears that describes multiple instances of the same document.

Documents in the document store can be combined into subsets to form new documents. This capability allows you to maintain a multi-level description of documents that are linked in hierarchical structures. Between documents, related hierarchical relationships are defined, subordinate relations that work in the direct (from the head document to the subordinate) and reverse (from the subordinate document to the head) directions.

Fig. 9 shows a fragment of a template for extracting information about a project from the website of a research project. This template allows you to extract such attributes of the Project class as "Title" and "Abstract", as well as arguments of the "Project Publication" and "Project Participant" relationships, i.e. objects describing publications about the project and project participants, respectively [13]. Each template designed to extract information is described by the Class block and contains blocks of attributes (Attr), relations (Relation), and arguments of relations (Object).

To extract information that makes up the context of the project and, as a rule, is determined by the relations of the Project class, for example, data on publications on the project topic, persons and organizations participating in the project, handlers and templates are used that are specially built to extract information of this type (Fig. 10) and are repeatedly used in other templates corresponding to such basic concepts of ontology as Publication, Person, Organization, etc. [14].

As mentioned above, information extracted from Internet resources is presented in the form of a semantic network of information objects, i.e. an oriented multigraph. The integration of the resulting graph into the EduDIS is performed by the information entry module.

To date, all the main components of this subsystem have been implemented and methods have been developed for extracting information about Projects, Personalities, Organizations, and Events, including related templates and handlers that implement information about publications.

Fig. 9. Extracting information from a data warehouse using a template

Fig. 10. Extracting information from a website using a template

6. Discussion of experimental results of building a model and technology of access to documents supporting scientific and educational activities

As the above overview of projects using the Z39.50 protocol and Z39.50 servers shows, document access technologies have an extensive history and many different ways of solving problems that arise during their implementation. Even though most technologies were developed in the interests of certain categories of users and are focused on meeting their information needs, the common features inherent in all document access technologies allow us to determine the basis for building a new access technology. This basis is documented search systems and the active use of metadata.

Thus, the organization of differentiation of user access to distributed information systems is a complex task, the solution of which consists of determining the basic requirements for the operation of systems in scientific and educational activities, taking into account additional requirements due to the distributed nature of information resources and the direct creation of developed infrastructure for the presentation and exchange of metadata. The proposed technology for supporting large information repositories is based on the client-server architecture of the IP and meets the above requirements.

For a research organization, such a system has a solid functional redundancy. The inclusion of access technology in the structure of the EduDIS organization used in the library limits the scope of such a system to bibliographic data only, making it difficult to use Z39.50 to solve problems of access to other data. In addition, the functional complexity of the system makes it difficult to test in case of technical problems and, as a result, it takes more time to eliminate them.

The use of a shared repository of documents and metadata and the document object model allows you to form a description of any document, whether it is a printed publication or an electronic document in any data schema, using the "native" DBMS used in the EduDIS library or XML language notation, creating prerequisites for the formation of an XML DBMS. The versatility of the data model used allows you to vary the software modules, creating the desired combinations.

The proposed approach to the creation of intelligent scientific and educational Internet resources is based on the technology of creating and maintaining an information environment of distributed learning. The information basis of EduDIS consists of an ontology, which, along with the traditional description of the subject area, includes a description of the structure and typology of the corresponding data warehouses and network resources. In addition, using the ontology framework, which is its declarative component, makes it easy to expand and configure the system so that both new knowledge and new sections of information resources can be connected to it.

Ontology offers effective means of presenting various information on a given topic, supports the systematization

and integration of relevant information resources, provides them with meaningful access.

According to the ontology, the following are automatically built:

– the scheme of the internal database (DB) of the IS (the logical structure of the database and its integrity constraints);

– forms for filling in the IS database with data (information objects that are instances of ontology concepts);

– the scheme of navigation through the information space of the IP (on the relations of ontology);

– forms of search queries (by concepts and relations of ontology).

Thanks to the use of ontologies as an information model, EduDIS is not only another catalog of resources on a given topic, but also a network of knowledge and data that allows you to maintain convenient navigation and search for content through your connections. The division of EduDIS ontology into subject-independent and subject-ontological adapts EduDIS to any field of scientific knowledge. The supposed possibility of declarative regulation of ontology when using EduDIS allows you to track the dynamics of the emergence of new knowledge and information resources on the topic and thereby support its relevance and usefulness.

The results of the work allow us to recommend the developed technology of access to documents for implementation in libraries and information centers of scientific and educational activities, as satisfying the requirements of users and providing access to diverse scientific publications. The technology makes it possible to implement information resources based on various conceptual approaches, as well as to use its software tools for promptly informing colleagues about the results of scientific research.

7. Conclusions

1. Thus, the model of information processes of document access technology can be represented as an interconnected set of functional modules. The division into modules is determined by the functionality and the software used in each of them. The use of a distributed repository of documents and metadata, as well as the document object model, makes it possible to form a description of a wide set of documents. The versatility of the data model used allows you to vary the software modules, creating the desired combinations.

A single access point implements the "single window" technology and is intended to serve as a user interface for accessing documents. The development and implementation of a document access point are carried out based on Web technologies. With the help of an access point, the user is notified about the databases available to him and can independently allocate a list of resources to search for the document he needs.

2. The results of the work allow us to recommend the developed technology of access to documents for implementation in libraries and information centers of scientific and educational activities, as satisfying the requirements of users and providing access to diverse scientific publications. The technology allows implementing information resources based on various conceptual approaches, as well as using its software tools for promptly informing colleagues about the results of scientific research.

Acknowledgments

References

1. Khamdamov, U., Abdullayev, A., Elov, J., Sultanov, D. (2020). Conceptual Model of the Education Management Information System for Higher Education Institutions. International Journal of Advanced Trends in Computer Science and Engineering, 9 (5), 7295–7300. doi: https://doi.org/10.30534/ijatcse/2020/59952020

2. Molnár, B., Benczúr, A. (2015). Modeling information systems from the viewpoint of active documents. Vietnam Journal of Computer Science, 2 (4), 229–241. doi: https://doi.org/10.1007/s40595-015-0046-9

3. Fedotov, A. M., Abdelieva, M. N., Baidavletov, A. T., Bapanov, A. A. Sambetbayeva, M. A., Fedotova, O. A. (2015). Conceptual model of scientific and educational information system. Vestnik Novosibirskogo gosudarstvennogo universiteta. Seriya: Informatsionnye tekhnologii, 13 (3), 52–67. Available at: https://cyberleninka.ru/article/n/kontseptualnaya-model-nauchno-obrazovatelnoy-informatsionnoy-sistemy

4. Sandugash, S., Jamalbek, T., Madina, S., Akbota, Y., Ainur, A. (2021). Building a Standard Model of an Information System for Working with Documents on Scientific and Educational Activities. International Journal of Advanced Computer Science and Applications, 12 (9). doi: https://doi.org/10.14569/ijacsa.2021.0120951

5. Podkorytova, N. I., Lakizo, I. G., Artemyeva, E. B. (2020). Scientific libraries in the structure of Novosibirsk Scientific-Educational Complex. Scientific and Technical Libraries, 1 (9), 61–77. doi: https://doi.org/10.33186/1027-3689-2020-9-61-77

6. Makeeva, O. V. (2020). New approach to building a competency model for library professionals. Proceedings of SPSTL SB RAS, 1, 71–77. doi: https://doi.org/10.20913/2618-7515-2020-1-71-77

7. Tsvetkova, V. A., Mokhnacheva, I. V., Rodionov, I. I. (2020). Scientific libraries in the information infrastructure of Russia. Vzaimovliyanie informatsionno-bibliotechnoy sredy i obschestvennyh nauk, 61–77. Available at: https://www.elibrary.ru/item.asp?id=44058881

8. Serikbayeva, S., Tussupov, J., Sambetbayeva, M., Muratova, G., Makhanov, M., Borankulova, G., Yerimbetova, A. (2021). Development of queries using the Z39.50 protocol in distributed information systems to support scientific and educational activities. Eastern-European Journal of Enterprise Technologies, 5 (2 (113)), 66–79. doi: https://doi.org/10.15587/1729-4061.2021.239540

9. Shokin, Y. I., Fedotov, A. M., Zhizhimov, O. L. (2015). Technologies for designing of distributed information systems to support research. Vychislitel'nye tekhnologii, 20 (5), 251–274. Available at: https://cyberleninka.ru/article/n/tehnologii-sozdaniya-raspredelennyh-informatsionnyh-sistem-dlya-podderzhki-nauchnyh-issledovaniy

10. Zhyzhymov, O. L., Fedotov, A. M., Fedotova, O. A. (2012). Building a generic model of information system for working with documents on the scientific heritage. Vestnik Novosibirskogo gosudarstvennogo universiteta. Seriya: Informatsionnye tekhnologii, 10 (3), 5–14. Available at: https://nsu.ru/xmlui/bitstream/handle/nsu/236/01.pdf?sequence=1&isAllowed=y

11. Fedotov, A. M., Barakhnin, V. B., Zhizhimov, O. L., Fedotova, O. A. (2014). A model of information system to support scientific and educational activities. Vestnik Novosibirskogo gosudarstvennogo universiteta. Seriya: Informatsionnye tekhnologii, 12 (1), 89–101. Available at: https://cyberleninka.ru/article/n/model-informatsionnoy-sistemy-dlya-podderzhki-nauchno-pedagogicheskoy-deyatelnosti

12. Sadirmekova Zh. B., Sambetbayeva, M. A. (2020). Development of technology for extracting metadata from documents during the integration of information systems. Science and Innovative Technologies, 1 (4), 209–214. doi: https://doi.org/10.33942/sit.nes004

13. Zagorulko, G. B. (2016). Development of ontology for intelligent scientific internet resource decision-making support in weakly formalized domains. Ontology of designing, 6 (4 (22)), 485–500. Available at: https://cyberleninka.ru/article/n/razrabotka-ontologii-dlya-internet-resursa-podderzhki-prinyatiya-resheniy-v-slaboformalizovannyh-oblastyah

14. Borovikova, O. I., Zagorulko, G. B., Zagorulko, Y. A., Shestakov, V. K. (2017). Use of patterns for the development of the ontology for the information-analytical internet resource "decision support". Informatsionnye i matematicheskie tekhnologii v nauke i upravlenii, 3 (7), 144–156. Available at: https://cyberleninka.ru/article/n/ispolzovanie-patternov-dlya-razrabotki-ontologii-informatsionno-analiticheskogo-internet-resursa-podderzhka-prinyatiya-resheniy