Building a Standard Model of an Information System for Working with Documents on Scientific and Educational Activities

Serikbayeva Sandugash[1], Tussupov Jamalbek[2]
Department of Information Systems, L.N. Gumilyov Eurasian National University, Nur-Sultan, Kazakhstan

Yerzhanova Akbota[4]
S. Seifullin Kazakh Agro Technical University
Nur-Sultan, Kazakhstan

Sambetbayeva Madina[3]
L.N. Gumilyov Eurasian National University, Institute of Information and Computational Technologies CSMES RK
Nur-Sultan, Almaty, Kazakhstan

Abduvalova Ainur[5]
NLC "Taraz Regional University Named after M.KH. Dulaty", Taraz
Kazakhstan

***Abstract*—To increase the effectiveness of research, it is necessary to have access to systematic information resources of scientific work. Therefore, in any field of science, it begins with research, the search for scientific information, but with the growing number of scientific articles, books, monographs, patents, the search for information becomes more and more difficult. Creating a unified information system that allows scientists to quickly get acquainted with the results of other scientific research and prevent their duplication. The article 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 that supports scientific and educational activities, the functional capabilities 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 provided. The paper analyzes the functional requirements for such systems. The paper describes a technological approach to creating a standard model of an information system to support scientific and educational activities organized in the form of an electronic library for working with documents on scientific heritage. The article describes the architecture of the information system and the principles of integration with the digital depository, the rules for the presentation and transformation of metadata.**

***Keywords*—Scientific and educational activities; distributed information systems; electronic library; metadata; model; search; interoperability; document; ontology; Z39. 50; SRU/SRW; Apache Solr**

I. Introduction

The rapid development of information technologies and means of data transmission has led to qualitative changes in the solution of one of the most important tasks facing humanity – the preservation of information for the purpose of its transmission. Until now, the main form of storing and distributing information was printed publications, and the main means of accessing information were libraries. New information technologies have provided new opportunities for solving the problems of creating repositories of information resources, their organization, means and ways of accessing them by users. In a generalized form, such approaches have become known as "Electronic Libraries (EL)". The information service on printed media has been replaced by the provision of users based on the electronic presentation of a wide variety of information, replicated in unlimited quantities and quickly accessible over global computer networks, regardless of the time of access to it and the location of users. The digital library should be understood more broadly as an environment that brings together "collections, services, and people to support the full life cycle of creating, distributing, using, and preserving data, information, and knowledge". The main tasks of the electronic library are the integration of information resources and effective navigation in them. The integration of information resources is understood as their integration in order to use (with the help of convenient and unified user interfaces – preferably one) different information while preserving its properties, presentation features, and user manipulation capabilities. However, the pooling of resources does not have to be physically performed. The main thing is that it should provide the user with the perception of the available information as a single information space. In particular, electronic libraries should provide work with heterogeneous databases or database systems, providing the user with the effectiveness of information searches, regardless of the features of the specific information systems to which access is made [1].

Distributed information systems to support scientific and educational activities operate with various types of information. These can be publications, electronic documents, electronic collections, ontological descriptions, data sets, logical descriptions, etc. These resources, which are in demand 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 search and identification.

Data integrated into an open semantic space is a collection of knowledge about a certain subject area. At the same time, the use of resources is accompanied by problems of determining the rules for obtaining up-to-date copies of resources and their transformation according to the specifics of a particular research environment. These problems are caused by the lack of adequate mechanisms to distribute the high-demand knowledge-intensive resources, as well as the heterogeneity of ways to represent and store resources in the form of files (binary, text, XML markup), spreadsheets, databases, electronic documents, catalogs, etc. Most often, these problems are solved in each individual case individually.

Special systems are needed to provide access to such resources. Therefore, providing universal ways to work with distributed and heterogeneous data, where it is not known in advance what types of objects the end user will have to work with, and unifying the presentation of this data, is the main task when integrating information resources in distributed information systems. At the same time, the idea of using Z39.50 technologies for resource integration looks very attractive, since to date, the Z39.50 standard is the only standard that regulates universal network access to databases based on an abstract data model [2].

The main requirement for information systems designed to support scientific and educational activities is interoperability.

The interoperability of any information system, including an electronic library, is understood as the degree of its ability to interact with other information systems, including with a person [3]. But if, when interacting with the latter (as with an information system), the main burden on ensuring mutual understanding falls on a person who is able to process even very poorly organized information, then special technological approaches and general agreements are required to ensure effective interaction between information systems. Ensuring the interoperability of systems is impossible without strict compliance with the relevant international standards and recommendations. At the same time, the standards must comply with:

- data access protocols and interfaces;
- search languages and interfaces;
- data representation schemes and formats;
- interfaces for visualizing the same type of data;
- rules for encoding information;
- data access control rules.

In [4], the basic profile of the information system standards for supporting scientific research, organized in the form of an electronic library, was defined.

The metadata profile refers to the adaptation of the existing metadata schema to the needs of a specific task being solved by the information system 1. Based on the analysis of the existing metadata formats intended for working with publications, documents and other information resources, it can be concluded that the most suitable format for research work with materials on scientific heritage is GOST 7.19-2001 (MEKOF). Compared to other commonly used metadata formats (formats of the MARC family), this format has the most complete classification system for document types and other information resources and a fairly large set of reference dictionaries necessary for describing and identifying information resources.

In this paper, we consider a technological approach to creating a standard model of an information system designed to support scientific research. The developed model of the information system for working with materials related to the scientific heritage should solve the problems of long-term storage of information, organization of abstract search by attributes, organization of collection and exchange of metadata and information between remote repositories of information resources.

In the information space, events, facts, and any other entities of the real world exist only in the form of documents [4]. As a result, the document is the main element of the system under consideration.

As you know, the information system (IS) and some parts of information systems are defined ambiguously in the scientific educational literature. Differences in definitions are usually caused by differences in the subject areas of application of the described IR and different approaches to the description of IR.

Let's consider a number of different information systems:

An information system is a set of information contained in databases and information technologies and technical means that ensure its processing [2].

An information system is an information processing system that works together with organizational resources, such as people, technical means, and financial resources that provide and distribute information [3].

An information system is a conceptual scheme, an information base, and an information processor that together make up a formal system for storing and manipulating information [4].

Information system: A system designed for storing, processing, searching, distributing, transmitting, and presenting information [5].

An information system is a material system that organizes, stores, and transforms information. This is a system in which the main subject and product of labor is information [6].

An information system is an applied software subsystem focused on the collection, storage, search and processing of textual and / or factual information [7].

Information system – a system designed for storing, processing, searching, distributing, transmitting and providing information [8].

A modern information system is a set of information technologies aimed at supporting the life cycle of information and including three main components of the process: data processing, management, information management and knowledge management [9].

Among Russian scientists in the field of computer science, the broadest definition of IR is given by M. R. Kogalovsky, in his opinion, an information system is a complex that includes computing and communication equipment, software, linguistic tools and information resources, as well as system personnel and provides support for a dynamic information model of some part of the real world to meet the information needs of users [10].

The part of the real world that is modeled by an information system is called its domain.

The dynamic model is understood as the variability of the model over time. This is a "live", working model, which displays the changes occurring in the subject area. Such a system must have a memory that allows it to store not only information about the current state of the subject area, but also, in some cases, the background.

Since the domain model supported by the information system materializes in the form of information resources organized in the necessary way, it is called an information model.

The above definition of M. R. Kogalovsky covers information systems of all types, in particular, fact-based systems that are based on database technologies and operate with structured data, text search systems that operate with documents in natural languages, the global hypermedia information system Web, etc.

In the textbook Information Systems, M. M. Telemtayeva considers IP as a complex and large system.

In relation to complex systems, it is based on the postulate of Academician A. I. Berg: "to create a model of a complex system, it is usually necessary to use more than two theories, more than two languages for describing the system, due to the qualitative difference in the internal nature of the system elements among themselves and the presence of different approaches to modeling objects of different nature"[11].

In relation to large systems, it is based on the definition given by V. I. Chernetsky: "a large system (BS) is a system that is a set of interconnected controlled subsystems united by a common control system, the characteristic feature of which is the presence of separate parts. Moreover, for each part, it is possible to determine: the purpose of functioning, subordinate to the general purpose of the entire system; the participation of people, machines and the natural environment in the system; the existence of internal material, energy and information connections between the parts of the system; as well as the presence of external links of the system under consideration with others" [12].

An information system is a complete set of parts that interact with each other and with the external environment of the IR. Parts of the IR are: database, information, information technologies, technical means, information products, models of internal and external user environments [13].

A system is a complete set of ways and / or means of ensuring the interaction of the internal environment of the elements (parts) of the system with the external environment of the system. The external environment of the system is usually structured from the point of view of the system in the form of sources of resources and consumers of the products of the system's activities;

Consistency is the integrity of an element (parts) of a system in relation to a given system. The element (parts) of the system is intended for activities in the interests of the system.

IR is a complete set of ways and means to ensure the interaction of the internal environment of the IR parts with the user of the IR-part of the external environment of the system that consumes the products of the IR activity.

An information system is an interconnected set of information, technical, software, mathematical, organizational, legal, ergonomic, linguistic, technological and other means, as well as personnel, designed to collect, process, store and issue information and make management decisions.

The information system consists of objects – elementary units of documents, and documents-information units. Many documents containing factual information, having the same physical structure and logical, informative purpose, form collections. Collections are characterized by their descriptions and descriptions of the structure of the documents that make up it [14].

A collection is a common form of organization of information resources that is determined by its parameters (style, attributes) and the structure of its documents and is a systematized collection of documents that are united by some criterion of belonging, for example, by content, purpose, access method, etc., provided with a meta description (metadata) in accordance with standards and data schemes. The document is characterized by its parameters (style, attributes) and the structure of the objects that it consists of. An object is defined by the type of data (according to the selected data schema) that it contains, and the description of the object's properties and methods [15].

Due to the fact that the information in the IR displays some entities of the real world (physical objects: objects, processes, phenomena, persons, publications, documents, algorithms, programs, files, facts, key terms, etc.), it is necessary to consider the IR as a set of information objects-data sets that represent (describe) Note that the development of the EB model should use ontological descriptions and conceptual models that summarize the accumulated experience in the field of creating and using EB [20]. A good overview of the existing conceptual models of EB is given in [22].

The ontological model of the IRIS EB is based on the conceptual models of the RM OAIS EB [23] and DELOS DLRM [24].

According to the DELOS conceptual model, an Information Resource (IR) is an abstract concept expressed by instances of one of its specializations. In particular, instances of the concept of IR are instances of an information object of any type (for example, documents, databases, collections, functions, etc.). Each resource in accordance with the DELOS model:

- has an identifier;
- organized according to the resource description. A resource can be complex and structured, because it, in turn, can consist of smaller resources and have connections to other resources;
- can be regulated by the functions that control its life cycle;
- expressed in terms of an information object;
- must be described with metadata, and can also be described or supplemented with additional metadata and annotations.

The implementation of the Electronic Library Management System is based on the meta-model, based on the fact that each information resource is characterized by a set of attributes inherent in it, and methods that characterize its properties and relationships with other resources. An effective means of describing information objects is metadata – data that is an integral part of an information object and describes a real object or group of objects.

Each information object in the IR consists of:

- Information content of the object (primary information object: for example, an image, full text, etc.) - an object that can be used independently;
- metadata object-an object whose main purpose is to provide information about the IR (usually about the primary information object);
- annotation object – an object whose main purpose is to annotate an IR or part of it. Examples of such annotations include notes, structured comments, and links. Annotation objects help to interpret the IR, contain either support, detailed explanations, or information about how the IR can be used.

II. Document Concept

One of the most important manifestations of human behavior is communication, i.e. communication with other people through certain signs or symbols. Initially, information about the surrounding world was transmitted by a person using gestures, facial expressions, shouting, touching, etc. - the simplest means of visual, auditory, tactile communication. The emergence of meaningful speech and language marked, according to a number of scientists, the emergence of the first information technology in the history of human society [16].

Meanwhile, with the development of man, the need to transmit information not only in space, but also in time, i.e., in the storage of information, increased. However, the simplest means of communication and information transmission were imperfect. The same human speech is heard only at a short distance and only at the moment of its utterance. It was difficult to retain the necessary information, since the knowledge was not yet separated from the subject who possessed it. It is no accident that at that time the role of a kind of knowledge banks and channels of their transmission was played by the elderly, i.e., the most experienced members of society.

The separation of information from the subject and the first attempts to consolidate it were associated with the use of signaling. To transmit information in space, signaling was used by smoke, fire fires, the sounds of trumpets, drumming, a branch or arrow placed in a certain way, etc. Objects were also used, which were given a symbolic meaning. The example of the symbolic message of the ancient Scythians to the Persians given by the ancient Greek historian Herodotus became a textbook. This message consisted of a bird, a frog, a mouse, and a bunch of arrows, and meant: "If you Persians do not learn to fly like birds, to jump through swamps like frogs, to hide in holes like mice, you will be showered with our arrows as soon as you enter the Scythian land."

Later, symbolic signaling was replaced by conditional signaling, in which objects were used as conditional signs by prior agreement of people about what a certain object would mean. As a result, there were systems of mnemonic signs for counting with the help of objects, as well as more complex "nodular writing": among the ancient Incas, in Ancient China, among the Mongols. Probably, this kind of "letter" was also available to the Slavs. It is no accident that the Russian language has preserved the expression "tie a knot for memory", i.e., to keep some information in memory. Tags (plaques) with notches were also used as conventional signs - in trade, financial, and creditor operations. Among the Slavs, such tags were called "noses", since they were usually carried with them, fixing any information with the help of various notches, notches. Hence, the expression “cut on the nose” is, remember it firmly [17].

Grave mounds, burial mounds, crosses, tombstones, property signs (heraldic signs, boundary stones, cattle brand marks, etc.) were used to consolidate and transmit information over time.

Objective ways of communication have been preserved to this day: the presentation of bread and salt as a sign of hospitality, bouquets of flowers and souvenirs as a sign of attention, military insignia, flags of states, traffic lights and semaphores, etc. The appearance of writing marked the transition of humanity to a new information technology. With the help of graphic sign systems, it became possible to separate information from the subject and fix it on some material for the purpose of subsequent transmission in time and space. As a result, there was documented information, i.e. a document.

The concept of "document" is currently the most common in the sciences that study different ways of storing and transmitting knowledge (or information) in society. There are many definitions of a document that have fixed ideas about it.

The term "document" comes from the Latin word "documentum". Documents appeared as an additional (to sound speech) means of communication of people. They were brought to life primarily by the need to capture, fix and transmit a particular message in time and space. Carriers, a material object on which information was recorded,

performed in ancient times mainly the function of evidence. Therefore, the Latin word "documentum" meant a sample, proof, testimony.

This term, in turn, came from the verb "docere" - to teach, to teach. The roots of this word go back to the Indo-European proto-language, where it meant a gesture of outstretched hands associated with receiving or receiving something. On the basis of this word, the words "doceo" were formed - I teach, I teach, "doctor" - a scientist, "doctrine" - teaching, and finally, "documentum" - what teaches an instructive example. In this sense, the word document was used, for example, by Caesar and Cicero.

Later, the word "document" acquired a legal meaning and came to mean "written evidence", "evidence drawn from books, supporting records, official acts". In the sense of a written certificate, the word "document" was used from the Middle Ages to the XIX century.

In the XIX century, a new aspect is highlighted: the importance of the document in management. For the designation of documents at this time, synonymous concepts were used: "business paper", "act", "case". The document was considered to be the information recorded in the form and intended for the implementation of the management process.

In the twentieth century, the term "business paper" is gradually replaced by «service document". In Soviet times, the term "document «was firmly established in normative acts, special literature, and work practice. It is still preserved at the present time.

Along with it, the concept of "act" is also used to refer to documents related to the field of management and law. They include almost all actions of the authorities and public administration, documented.

Over time, the term "historical documents" appears. They are considered chronicles, chronicles, notes, and other written sources that indicate a historical event, person, epoch, etc.

Thus, in the definition of the document, three aspects can be distinguished - legal, managerial, and historical.

The word "document" came to the Russian language in the time of Peter I, as a loan from the German and Polish languages, in the meaning of a written certificate [18]. At the beginning of the XX century, it had two meanings:

1) any paper drawn up in a lawful manner and can serve as proof of rights to something (property, fortune, free residence) or to perform any duties (conditions, contracts, debt obligations);

2) in general, any written evidence.

By the second half of the XIX century, the terms derived from the word "document" appeared in reference publications of some countries of the world: documentation-in the meaning of the preparation and use of documented evidence and authority; documentary - related to the document.

At the end of the XIX century, there is a tendency to narrow the boundaries of the concept of "document": first it was considered as any object that serves to obtain and prove, then - as a written certificate confirming certain legal relations. The concept was used mainly in the legal sense.

Since the beginning of the XX century, a new, broader understanding of the concept of "document" has been introduced into the term system: it was introduced by a well - known Belgian scientist, the founder of documentation - the science of the totality of documents and the field of practical activity-Paul Hautelet-in his treatise on Documentation, defines the concept (term) Document: "a material object containing information, specially designed for its transmission in space and time" [19] - which is interpreted as the main "object" with which any information system operates[20]. Thus, a document is an information object that represents a structured description of a real entity (object, subject, fact or concept), the totality of which makes up the information content of the system. The document presented in electronic form has a certain standard set of attributes and allows for unambiguous identification. A document can describe an article from a journal, the journal itself, a person, a digitized image, experimental data, a program or computational algorithm, a database, a fragment of a database, etc.

P. Otle first used a comprehensive approach to the typological classification of documents, taking into account the content and form of the document, in the "Treatise on Documentation".

The scientist divided the entire set of documents into three main classes:

1) Bibliographic documents, i.e., texts that are traditionally considered works of writing and printing. Among them are brochures, monographs, essays, treatises, manuals, encyclopedias, dictionaries, periodicals and continuing publications (magazines, newspapers, yearbooks, etc.). In addition to these, bibliographic documents included texts of personal origin (letters), official messages and accounting (registration) books (or magazines), as well as signs, slogans, tickets and other travel documents.

2) Other graphic documents, i.e., non-text documents: cartographic, pictorial, musical notation. Among the pictorial ones are: iconographic, containing a printed image (prints, engravings, postcards, etc.); photographs; documents perceived through projection devices (including microcopies). As a special variety, "pictorial monuments are distinguished: inscriptions, coins, medals, seals (stamps).

3) Documents - substitutes for books: discs, phonograms, films, and along with this-radio broadcasting (recording and transmitting sound), television, including telephotography, radio telephotography and television itself.

A special place in this classification series was occupied by" documents of three dimensions": natural (minerals, plants, animals) and artificial, created by man (materials, products, technical objects, as well as medals, models, reliefs). They also include scientific tools, didactic materials, and visual aids. Especially highlighted among them are three-dimensional works of art: works of architecture and sculpture.

It is in this broad sense that the concept of "document" was later used when it comes not only to the collections of libraries, archives, museums, information services, but also to social, in particular, mass communications in general.

We can distinguish the following values of "document", introduced by P. Otle:

1) Any source of information, transmission of human thought, knowledge, regardless of whether it is embodied in a material-fixed form or is a conductor (transmitter) of information in time, can be considered a document. This concept covers material objects-information carriers, as well as, radio, television, and theatrical performances.

2) Documents are material objects with recorded information collected by a person to create any collections. This includes both artificial objects created by man, and natural, technical objects located in the museum.

3) Documents also include material objects created by a person specifically for recording, storing and reproducing information in order to transmit it in space and time, regardless of the method of recording. These are both "written" documents (i.e., with information recorded by writing characters), and visual, phonographic recordings and films (the results of machine recording of images and sound).

The author of the" Treatise... "repeatedly emphasizes the synonymy of the concepts of" document "and" book"; from the context, it can be understood that he considers the former as broader.

Thus, P. Otle entered the world history of documentary studies as the founder of documentation-science and practice. He was the first not only to introduce the basic concept of "document" into scientific use, but also to reveal its broadest meaning. P. Otle made the first attempt at a comprehensive classification of documents by a set of features. Although it had significant drawbacks, the author managed to group the existing variety of information sources that function in social communication. Subsequently, the theoretical thoughts of domestic and foreign specialists moved in the same direction as the thought of the documentarian P. Otle.

The concept of P. Otle considers the document as a carrier of social information. However, in reference publications of that time, there continues to be a narrow meaning of this word: in addition to the legal one, the concept of "historical document" (a fixed certificate of an era, person, etc.) and "accounting document" (serving as the basis for carrying out economic actions-receiving and issuing valuables) is introduced.

Paul Marie Ghislain Otlet is a Belgian writer, entrepreneur, thinker, bibliographer, lawyer, and peace activist. He was one of those who are considered the founding fathers of computer science. He is the author of numerous publications on the problems of book studies, bibliography and documentation. He was an active promoter of international cooperation in the field of book studies.

In 1934, he published his famous book "Trait de documentation", which laid the foundation for the science of documents (in the broad sense), which can now be called "documentation" - the forerunner of modern computer science, the author of the concept of the information universe (electronic, but not digital), as the development of telephone communication and television. They are also the developers of the Universal Decimal Classification (UDC) - one of the most outstanding examples of faceted classification.

P. Otle repeatedly emphasizes the synonymy of the concepts of "document "and" book"; from the context, it can be understood that he considers the former as broader. He was the first not only to introduce the basic concept of "document" into scientific use, but also to reveal its broadest meaning. P. Otle made the first attempt at a comprehensive classification of documents by a set of features. The concept of P. Otle considers the document as a carrier of social information.

Consider a number of other document definitions:

A follower of P. Hautelet, Suzanne Brie (1894-1989), in 1951 published a work entitled "What is documentation?", which begins with the following statement: "A document is a certificate confirming a certain fact", "it is any physical or symbolic sign, preserved or recorded, intended to represent, reconstruct, confirm some physical or individual phenomenon". Suzanne Brieux equated the document with an organized physical observable object. This approach resembles the definition of «material culture". in terms of cultural anthropology and the ideological (more precisely, methodological) approach of "object as a sign" in semiotics. Also in the 1951 Manifesto on the Nature of Documentation, it stated: "A document is evidence in support of a fact ... it is any physical or symbolic sign, stored or recorded, intended to represent, recreate, or demonstrate a physical or conceptual phenomenon. " [21].

According to S. Brie, the wild antelope is not a document. But if it is placed in a zoo cage and studied, it becomes a physical object, a primary document, and all the articles about it are secondary, derived documents. (Here we are talking about the antelope, which is a new, newly discovered species of African antelope).

The definition of a document is proposed by Yu. N. Stolyarov "Document - semantic information created by a person specifically for social communication and recorded in any way on any medium" [22]. Where "fixing information" defines as an essential characteristic of a document: "A document is an object that allows you to get the required information from it" [23]. He also states, "The document status can have any object", "the same object may or may not be a document - it all depends on whether it serves to get information from it (from it, from it) or not." Any objects of reality can provide information, but not all of them become documents. For example, "Madame Brieux's antelope" can become a document if it is recorded in the form of an image (photo, drawing, sculpture, etc.) or if the words "Madame Brieux's antelope" are recorded, and the real antelope, even if it has such an original name, is not a document. That is, "A document is an object that allows you to get the required information from it."

A. V. Sokolov's definition: "A document is a stable material object intended for use in social semantic communication as a completed message". Where the author draws attention to the distinctive features of the document: the presence of semantic content, stable material form, intended for use in communication channels, the completeness of the message. Note that the presence of semantic communication is the essence of any information process. The sign of the stability of existence is a mandatory characteristic of the document as a way of storing information.

III. Models of Distributed Information Systems to Support Scientific and Educational Activities

The rapid development of global information and computing networks leads to a change in the fundamental paradigms of data processing, which can be described as a transition to the support and development of distributed information resources [21]. Therefore, the most important task associated with the technology of working with information is to study ways to integrate distributed data sources.

Integration of information resources is understood as combining them in order to use (using convenient and unified user interfaces – preferably one) different information while preserving its properties, presentation features, and user manipulation capabilities. However, the pooling of resources does not have to be physically performed. The main thing is that it should provide the user with the perception of the available information as a single information space. In particular, electronic libraries should provide work with heterogeneous databases or database systems, providing the user with the effectiveness of information searches, regardless of the features of the specific information systems to which access is made.

An urgent task is to create a model of a distributed information system to support scientific and educational activities:

- to unify the process of sharing the results of scientific research;
- operate with data and documents integrated into an open.
- semantic space;
- provide services for the transformation of heterogeneous resources that implement the means of description, representation, automatic linking of resources, as well as interaction with search and classification mechanisms in accordance with the needs of users.

The model should provide the following functionality:

- publishing resources, including registration, naming, annotation, and format definition procedures;
- analytical processing of resources;
- access to published resources, including dynamic generation functions;
- for automated operation, you need a function for monitoring resources and updating their meta descriptions, functions for notifying users about the appearance of new resources and updating existing ones, and a dispatching function [24].

IV. Metadata Profile in Distributed Information Systems to Support Scientific and Educational Activities

An effective means of describing information objects is metadata, which is an integral part of an information object and describes a real object or group of objects.

An important property of metadata is its specificity with respect to the scope of the described objects (resources). Metadata can characterize entities that relate to both the virtual (information) space and the real world (persons, organizations, events). Metadata can be part of information resources, or it can be stored separately from information resources.

Metadata is necessary for solving the following tasks:

1) providing information about documents, their content, structure, methods of use, etc.;
2) systematization and classification of documents;
3) organization of in-system processing procedures;
4) Support for sharing with external IS.

Standards application profiles are created for a specific group of functional tasks or users. This makes it easier to create systems that work with metadata. A profile can be defined as "one or a combination of several basic standards with the identification of the selected classes, subsets, optional capabilities, and parameters of these basic standards required to perform a specific function" [22].

In the metadata area of resources for publications, the profile should contain a list of mandatory elements present in the resource description, and set dictionaries for describing the values of elements that complement or extend the acceptable set of values defined in the standard. In addition, additional description elements may be suggested.

Thus, the basis for the development of a scientific and educational system consists of standards and international recommendations that form the profile of a scientific and educational system, which is understood as one or a set of several basic normative and technical documents (standards and specifications) aimed at solving a specific task (the implementation of a given function or a group of functions of an application or environment), indicating, if necessary, the selected classes, subsets, options of basic standards necessary to perform a specific function [27]. The most important is the metadata profile of the information circulating in the system.

V. Requirements for Distributed Information Systems for Scientific and Educational Activities

Distributed information resources the development of the organization's information resources leads to the need to create an infrastructure for their integration into a single

information system that provides transparent access to distributed information.

The development of global information and computing networks today leads to a change in the fundamental paradigms of working with information resources. Today, the transition to distributed resources, the creation of infrastructure for and integration into a single information system that provides transparent access to distributed information is relevant.

Therefore, the most important task related to information technology is to research ways to integrate distributed data sources and create scientific groundwork in the field of distributed information systems and databases in order to develop technology that supports the creation and operation of large-scale information infrastructures based on virtual integration. This technology will allow you to create global infrastructures from dozens and hundreds of heterogeneous databases and solve strategic tasks in the field of automation of various forms of distributed activities. A narrower goal is to develop principles and software tools for virtual integration of distributed data sources based on international standards and recommendations for creating large-scale information infrastructures designed to virtualize data access to various DBMS using common rules and policies [23].

The tasks of distributed, as well as conventional, information systems are to store information and provide it to users in a convenient form. As a rule, such systems can be organized on the basis of various technological solutions aimed at implementing a particular distribution paradigm. Based on the main functions of information systems, various aspects of distribution can be considered:

1) Distributed information storage (distributed storage, network storage systems, and network file systems).
2) Distributed DBMS (adding, upgrading, changing data).
3) Distributed information access management and distributed information management.
4) Search for information in distributed sources.
5) Extracting information from distributed sources.
6) Visualization of information from distributed (heterogeneous) sources in unified user interfaces [24].

Distributed information systems represent an increasingly important trend for computer users. Distributed processing is a method for implementing a single logical set of processing functions on multiple physical devices, so that each performs some part of the overall required processing. Distributed processing is often accompanied by the formation of a distributed database. A distributed database exists when data items stored in multiple locations are interconnected, or if a process (program execution) in one location requires access to data stored elsewhere.

Distributed information systems to support scientific and educational activities is designed to collect, classify, analyze text publications of the Kazakhstan segment of electronic mass media for the management of information resources.

The purpose of creating the system: To develop a system for distributed information systems to support scientific and educational activities, to create a program to explore the capabilities of the Apache Solr platform for processing distributed data that uses big data technologies.

A set of the most general functional requirements for the IP support of scientific and educational activities was identified:

1) Collection of information resources.
2) Relevance of documents.
3) The relevance, completeness, and authenticity of the origin of the documents.
4) Use of intelligent services for processing user requests.
5) Knowledge extraction.
6) Support for non-centralized information system architectures.
7) Structuring of the information space.
8) Adaptive presentation of information.
9) Historicity of information.
10) Archive.

In the conditions of working in a distributed environment, the requirements for the IP support of scientific and educational activities are:

- support for the adopted metadata standards for data export and import;
- support for information exchange protocols with other information systems;
- support for the ability to link to internal resources both in user interfaces and at the system level.

System tasks:

1) Collection, storage and selection of unique publications from the Internet space to the system database.
2) Distribution of publications by topic: clustering, classification, definition of thematic combinations, ranking and filtering (by social spheres, regions, industries, etc.)
3) Definition of information occasions.
4) Calculation of the degrees of informative features of the publication, such as: collective use of purchased electronic literature catalogs, databases and bibliographic publications.
5) Definition of information trends.

VI. Architecture of Distributed Information Systems to Support Scientific and Educational Activities

Our generation has a rich scientific heritage that should be preserved.

Delay in this work can lead to irreparable losses associated with a temporary factor: the loss of documents, the departure from the lives of eyewitnesses to the events. One of the most important tasks related to the preservation of scientific heritage is a set of measures aimed at creating specialized information systems (electronic libraries) designed to store information, to organize access and mechanisms for using information [25].

Two necessary requirements that can be imposed on such information systems are obvious. The first requirement is the need to create and provide a system for reliable long-term storage of digital (electronic) documents while preserving all the semantic and functional characteristics of the original documents. The second is to provide a "transparent" search and access to users ' documents, both for review and for analysis and scientific work [23].

In the existing developments of electronic libraries, as a rule, the search and access to information is provided only through visual graphical interfaces. This is good for the human user, but very bad for the application user (for example, for conducting various analytical studies).

To provide search functions outside of graphical interfaces, support for special network services and query languages is required. Ideally, all information systems should support a single search profile and a single query language. 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 [5; 6] and the simpler SRW/SRU model [6]. The practical implementation of services such as SRW / SRU provides a significantly new quality of the electronic library – the ability to include its resources in global search engines at a higher level than the level of external indexing of static Web pages by other systems. Other possible search types are related to search by specified templates and to search using ontology. The search involving ontology is more intelligent. Its implementation requires additional information about the domain, including definitions of terms, entities, and relationships. It should be noted that the presentation of this additional information must comply with global agreements and international standards, otherwise the search using dictionaries, thesauruses and ontologies will always be limited to the current system, and interoperability will not be implemented [26].

As part of the tasks set, an information system architecture was developed (Fig. 1) to systematize the resources of the electronic library, a multi-level EC architecture is used, consisting of a data warehouse, a repository, a metadata server, an application server, a dictionary, reference books, as well as a software implementation of the developed architecture, deployed on existing hardware and put into operation.

Based on the formulated requirements, the information system for supporting research on scientific and educational activities should consist of a long-term storage system and an information management system for organizing an abstract search necessary for the analysis and conduct of scientific works. A very important component of the technology of working with scientific heritage is metadata, which contains the information necessary to document the process of storing information resources. This metadata is information about the format, structure, and use of information resources, the history of all operations, including any changes, authenticity, technical history, responsibilities, rights, and so on.

Thus, the information system on scientific heritage should functionally consist of three blocks.

1) Digital Depository 6 (or repository, hereinafter referred to as DD) is an independent system of long-term storage and access to heterogeneous digital objects, which is designed to provide electronic (digital) versions of documents on the scientific heritage (books, scientific articles, reprints, letters, images and other materials presented in electronic form).

2) Reference books are a set of databases containing information about authors and other persons (authoritative files), geographical locations, cities, publishing houses related to a particular scientific school, thematic dictionaries-classifiers, thesauri, descriptions of the subject area of this scientific school and classifiers of documents in accordance with the IECOF.

3) The metadata server should provide metadata management-cataloging of all information resources in accordance with generally accepted international standards. It should run a whole set of application services that should: support abstract search schemes in accordance with the schemes proposed by the Z39.50 protocol and SRW/SRU, support search schemes based on specified templates and using ontology, support fact detection and document identification based on information that is in the directory, as well as provide metadata collection from its own and remote data centers (exchange, synchronization and modification), metadata conversion between existing standards (GOST, MARC, etc. and the corresponding translation of metadata schemas from one format to another.

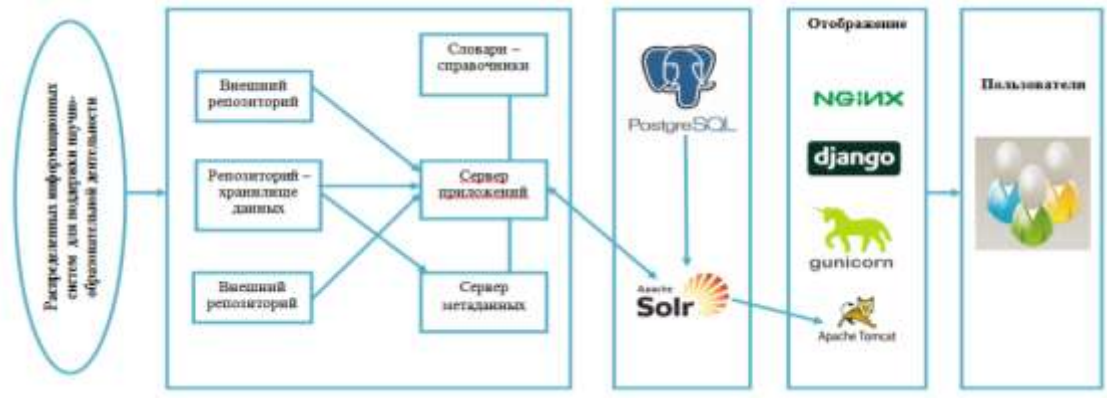

Fig. 1. Architecture of Distributed Information Systems to Support Scientific and Educational Activities.

Accurate categorization of the material using the dictionary of the reference book increases the probability that the search results will find documents relevant to the search expression when organizing a search in one or more electronic libraries [27].

4) PostgreSQL-acts as a persistent storage for structured data. The main types of data stored in this database:

a) News and metadata;

b) Processed data at the level of different basic units of analysis (token/word/phrase/sentence/text), including vectorization, lemmatization results, cleaning, etc.

c) Results of thematic modelling;

Results of news classification by various criteria (tone, politicization, social significance, etc.).

5) Apache Solr is a popular, fast-growing open-source search platform built on Apache Lucene. Solr is highly reliable, scalable, and fault tolerant, providing distributed indexing, replication, and load balancing, automatic disaster recovery, centralized configuration, and more. Solr supports the search and navigation functions of many of the largest Internet sites in the world. Since Solr has distributed search and replication capabilities, Solr is highly scalable. Here are some of the main features that solr provides:

a) Advanced full-text search capabilities

b) Optimized for high volume traffic

c) Open interfaces based on standards-XML, JSON and HTTP

d) Comprehensive administration interfaces.

e) Easy monitoring.

f) High scalability and fault tolerance.

g) Flexible and adaptable with simple configuration.

h) Next to the real-time index.

i) Extensible plugin architecture.

Data Processing:

When developing the architecture for data processing, the following main needs were identified:

1) The ability to parallelize calculations, including on multiple machines;

2) Flexible scheduling of various data processing tasks;

3) The ability to monitor the execution of tasks in real time, including prompt notification of exceptions;

4) Flexibility in the tools and technologies used.

Distributed information systems will contain the following subsystems:

- Subsystem-a repository of digital objects that provides user and administrative WEB - interfaces for accessing digital objects and collections, as well as interfaces for integration with other subsystems based on open international standards.
- A subsystem for managing current research information (SUEB), which includes information about the publications of employees, their participation in conferences and in the implementation of research projects.
- The subsystem will include user and administrative interfaces, as well as interfaces for integration with other subsystems based on open international standards.
- Subsystem for integration of distributed information resources based on Apache Solr technologies.
- Subsystem for access to distributed information resources based on technologies-Nginx, Djang, Apache Tomcat.

These subsystems together should provide:

- identification of information resources;
- identification, authentication and authorization of users;
- metadata management;
- information resource management;
- collecting statistics;
- monitoring the availability of services and resources.

In distributed search, a collection is a logical index on multiple servers. The part of each server that runs the collection is called the core. Thus, in an unallocated search, the core and the collection are the same, since there is only one server.

VII. Conclusion

Based on the formulated requirements, a prototype of an information system has been developed that can be used as a standard for working with documents in the field of scientific and educational activities, since it solves the main tasks facing these systems: ensuring reliable long-term storage of digital (electronic) documents while preserving all the semantic and functional characteristics of the source documents; ensuring "transparent" search and user access to documents for both familiarization and analysis of the facts contained in them; organizing the collection of information on remote digital repositories.

The developed model of the information system can be used as a standard model of the system for working with documents related to scientific heritage, since it solves the main tasks required for these systems:

- providing a system of reliable long-term storage of digital (electronic) documents with the preservation of all semantic and functional characteristics of the original documents;
- providing a "transparent" search and access to users ' documents both for review and for analysis and scientific work.

To date, all the components necessary to create a qualitatively new scientific information system are available and worked out clearly and systematically. Most of the scientific-centralized distributed systems allow you to create a single environment for the exchange of scientific information.

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] Zhizhimov O. L., Fedotov A.M. Ensuring the interoperability of electronic libraries / / Information technologies and mathematical modeling in science, technology and education (Bishkek, Kyrgyzstan, 5-9 October 2011): Izv. Kyrgyz State Technical University. Razzakov State University. 2011. No. 24. Materials of the international conference. pp. 331-335.

[2] Larkov N. S. Documentovedenie: Uchebnoe posobie / N. S. Larkov. - M.: AST: Vostok-Zapad, 2006.

[3] Larkov N. S. Documentovedenie: Electronic textbook. Tomsk: TSU 2002.

[4] Otle P. Biblioteka, bibliografiya, dokumentatsiya: Izbrannye trudy pionera informatiki [Library, bibliography, documentation: Selected Works of the pioneer of Informatics]. - Moscow: FAIR-PRESS: Pashkov House, 2004. - 348, [1] p.- (Special publishing project for libraries). - Bibliogr.: pp. 312-327. - Names. decree: pp. 340-342. - ISBN 5-8183-06.

[5] Fedotov A. M. Metodologii stroeniya razdelennykh sistem [Methodologies for constructing distributed systems]. 2006, vol. 11, Selected reports of the X Russian Conference. Distributed information and computing resources. (DICR-2005), Novosibirsk, October 6-8, 2005, pp. 3-16.

[6] Fedotov A.M., Zhizhimov O. L., Fedotova O. A., Barakhnin V. B. Model of information system for support of scientific and pedagogical activity // Vestn. Novosibirsk State University. Ser. Inform. technologies. 2014. Vol. 12, no. 1. pp. 89-101.

[7] Fedotov A.M., Barakhnin V. B., Zhizhimov O. L., Fedotova O. A. Technology of creating corporate information systems for accounting of scientific workers ' works. Novosibirsk State University. Ser. Inform. technologies. 2011. Vol. 9, issue. 2. pp. 31-41.

[8] Zhizhimov O. L., Fedotov A.M., Fedotova O. A. Building a typical model of an information system for working with documents on scientific heritage. Novosibirsk State University. Ser. Inform. technologies. 2012. Vol. 10, no. 3. pp. 5-14.

[9] Kogalovsky M. R. Metadata, their properties, functions, classification and means of representation // Proceedings of the 14th All-Russian Scientific Conference " Electronic Libraries: Promising Methods and Technologies, Electronic Collections — - RCDL2012, Pereslavl-Zalessky, Russia, October 15-18, 2012.

[10] Kogalovsky M. R. Metadata in computer systems/M. R. Kogalovsky / / Programming, 2013, N # 4. - p. 28-46.

[11] Kogalovsky M. R. Scientific collections of information resources in electronic libraries. Proceedings of the First All-Russian Scientific Conference "Electronic Libraries: Promising Methods and Technologies, Electronic Collections", St. Petersburg, October 1999. St. Petersburg University Press, 1999.

[12] Bezdushny A. N., Bezdushny A. A., Serebryakov V. A., Filippov V. I. Integration of metadata of the Unified Scientific Information Space of the Russian Academy of Sciences. Moscow: Raschut. A. A. Dorodnitsyn Center of the Russian Academy of Sciences, 2006. 258 p.

[13] Functional requirements for bibliographic records: conceptual model: graduate. report / translated from English by V. V. Arefyev. Moscow: Russian State Library,2006. 150 p. Shokin Yu. I., Fedotov A.M., Barakhnin V. B. Problems of information search. Novosibirsk: Nauka, 2010. 198 p.

[14] Kulagin M. V., Lopatenko A. S. Scientific information systems and electronic libraries. The need for integration. // Collection of proceedings of the Third All-Ross. conf. on electronic.libraries. RCDL ' 2001 Petrozavodsk, September 11-13, 2001, pp. 14-19.

[15] Fedotov A. M. Metodologii stroeniya razdelennykh sistem [Methodologies for constructing distributed systems]. 2006, vol. 11, Selected reports of the X Russian Conference. Distributed information and computing resources. (DICR-2005), Novosibirsk, October 6-8, 2005, pp. 3-16.

[16] Fedotov A.M., Zhizhimov O. L., Fedotova O. A., Barakhnin V. B. Model of information system for support of scientific and pedagogical activity // Vestn. Novosibirsk State University. Ser. Inform. technologies. 2014. Vol. 12, no. 1. pp. 89-101.

[17] Fedotov A.M., Barakhnin V. B., Zhizhimov O. L., Fedotova O. A. Technology of creating corporate information systems for accounting of scientific workers ' works. Novosibirsk State University. Ser. Inform. technologies. 2011. Vol. 9, issue. 2. pp. 31-41.

[18] Zhizhimov O. L., Fedotov A.M., Fedotova O. A. Building a typical model of an information system for working with documents on scientific heritage. Novosibirsk State University. Ser. Inform. technologies. 2012. Vol. 10, no. 3. pp. 5-14.

[19] Kogalovsky M. R. Metadata, their properties, functions, classification and means of representation // Proceedings of the 14th All-Russian Scientific Conference " Electronic Libraries: Promising Methods and Technologies, Electronic Collections — - RCDL2012, Pereslavl-Zalessky, Russia, October 15-18, 2012.

[20] Bezdushny A. N., Bezdushny A. A., Serebryakov V. A., Filippov V. I. Integration of metadata of the Unified Scientific Information Space of the Russian Academy of Sciences. Moscow: Raschut. A. A. Dorodnitsyn Center of the Russian Academy of Sciences, 2006. 258 p.

[21] Functional requirements for bibliographic records: conceptual model: graduate. report / translated from English by V. V. Arefyev. Moscow: Russian State Library,2006. 150 p. Shokin Yu. I., Fedotov A.M., Barakhnin V. B. Problems of information search. Novosibirsk: Nauka, 2010. 198 p.

[22] Kulagin M. V., Lopatenko A. S. Scientific information systems and electronic libraries. The need for integration. // Collection of proceedings of the Third All-Ross. Conf. on electronic.libraries. RCDL ' 2001 Petrozavodsk, September 11-13, 2001, pp. 14-19.

[23] S.K.Serikbayeva , D.A.Tussupov, M.A.Sambetbayeva, A.S. Yerimbetova, Taszhurekova ZH.K., Borankulova G.S., EduDIS construction technology based on Z39.50 protocol / / Journal of Theoretical and Applied Information Technology. - 2021. - Vol.99(10), -pp. 2244–2255.

[24] Serikbayeva S.K, Batyrhanov A.G., Sambetbayeva M.A., Sadirmekova Zh.B., Yerimbetova A.S. Development of technology to support large information storage and organization of reduced user access to this information: (IJACSA) International Journal of Advanced Computer Science and Applications, 2021 г.. - 7 : Т. 12. - стр. 493-503.

[25] 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.

[26] 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.

[27] 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.