The Creation of Information Model of Digital Library for Supporting Scientific and Educational Activity

A. Bapanov[1], J. Tussupov[1], A.Fedotov[2] and O. Fedotova[3]

***Abstract*—The work is dedicated to the creation of information model of digital library for supporting scientific and educational activity. The information needs of modern users and data objects, which describe basic entities of the scientific information space such as the publication of document, person, dictionary entry, function and user, also the relationships between them are discussed.**

Keywords—conceptual model; digital library; scientific and educational activities.

I. INTRODUCTION

Modern information technologies significantly affect practically all stages of scientific and educational process, which affects the changing information needs of students, scientists and teachers. The modern learner, armed with a computer, cannot be satisfied with the traditional mode of the educational process and the usual formats of teaching materials, whether textbooks, books or even simple text files. Besides providing students with training materials, it is also necessary to provide them with various search and classification services. The systematization and classification of available information resources in accordance with the needs of user is one of most important tasks of supporting both scientific and educational activities [1, 2].

Digital Library (DL) - the structured and cataloged collections management system of diverse digital documents. DL provides navigation and search tools. DL is able to provide not only a multilateral search and navigation in catalogs, but also to provide the user with a directly found resource (publication, document, photo, fact description, etc.), as well as additional information about it, for example, information about authors, bibliography, organizations and etc [3].

[1]Department of Information Systems, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
[2]Institute of Computational Technologies SB RAS, Novosibirsk, Russian Federation
[3]State Public Scientific Technological Library SB RAS, Novosibirsk, Russian Federation

Currently, there are many DL systems, but we can say with confidence that there is no universal system among them that could meet all the requirements of users. The analysis of existing DL systems shows their diversity at the levels:

- the information model that they provide;
- supporting users and group of users;
- functionality;

A number of problems arise because of that diversity and ignoring the needs of users:

- the integration of information from various DL;
- comparison of DL by provided functionality;
- evaluation and comparison of the performance of various DL systems;
- adding new types of stored objects;
- adding new functionality;
- backup, etc. [4].

For solving these problems, it is necessary to use conceptual models that generalize the gained experience in the sphere of creation and use of DL.

II. THE EXISTING CONCEPTUAL MODELS

Currently, a number of models are developed in the world that describe information resources in the scientific sphere, in the field of digital libraries, cultural and scientific heritage. The most interesting are the reference models, as they provide abstract view of the entities and of relationships that make up the content of the system.

The reference model is an abstract view of concepts and relations between them in some problem area. On the basis of reference model, constructed more concrete and detailed models, eventually embodied in real objects and mechanisms.

We list the most popular of them:

DL RM (Digital Library Reference Model) - a reference model of the digital library developed by a group of specialists of the association in the field of DL DELOS in 2006-2007, based on the analysis of existing library systems.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

The advantages of DL RM should include the greatest completeness of coverage among the existing conceptual models of DL. Also, much attention is paid to the functionality of the digital library in the model.

CIDOC CRM (Committee on Documentation Conceptual Reference Model) - is a formal ontology designed to describe information in the field of cultural heritage. The development of the model began in 1996, and in 2006 it became the standard (new version of ISO 21127: 2014) - A reference ontology for the interchange of cultural heritage in formation.

OAIS RM (Open Archival Information System Reference Model). This model has been used by many organizations to develop sets of metadata and organize large repositories of digital objects. On the basis of this model, was created the concept of an "institutional repository" as a system of long-term storage, accumulation of information and providing reliable access to digital objects, which is the result of intellectual activity of a scientific or educational institution [5].

FRBR (Functional Requirements for Bibliographic Records) - development by the International Federation of Library Associations and Institutions (IFLA) - ER-model of a documentary system working with bibliographic information.

CERIF - (Common European Research Information Format). The CERIF model is designed to support the management of research information, as well as the settings and interactions between research information systems and scientific digital libraries.

ESIS (ISIR) of the Russian Academy of Sciences (RAS) - Unified Scientific Information Space (Integrated System of Information Resources) presents a model of documentary IS [6], implemented as a control system for digital libraries. Distinguished four main groups of entities: participants in scientific activity, scientific activity, results of scientific activity, documents and publications.

IDIS Siberian Branch of RAS - Integrated distributed information system [3] represents a model of documentary DL implemented as a control system for digital libraries. Basic entities: document, person, organization, event, fact.

As a basis for the developed conceptual model taken models such as DL RM and OAIS RM.

III. THE CONCEPTUAL MODEL

The conceptual model of the digital library (DL) should describe what entities can be represented in the DL, and also should fix the rules and connections, which in particular involves the classification of entities, abstraction, generalization. The main goal of developing a conceptual model of a digital library for support scientific and educational activities is developing a model with a specific architecture for its subsequent implementation in the form of a software system.

The Information Resource is the most general concept, including any essence of the digital library. In the information space all information resources: documents, publications, persons, events, facts, programs and any other entities of the real or virtual world - exist only in the form of some information objects. Each resource in accordance with the reference model (for example, DL RM or OAIS RM):

- has an identifier;
- organized in accordance with the description of the resource. A resource can be complex and structured, because it, in turn, can consist of less resources and have connections to other resources. From an organizational point of view, resources can be grouped into sets of resources that are treated as a single entity;
- can be regulated by functions that control its life cycle, is characterized by a set of inherent attributes and methods that characterize its properties and connections with other resources;
- expressed through an information object.

In the digital library, each Resource corresponds to an Information object, which is a traditional secondary information object that contains a description of the primary resource, i.e. an information object is an object that stores information about objects of DL (physical objects, resources, information objects).

The information object is the most general concept in the system, representing an arbitrary unit of information in the DL.

Since the information object is a resource, it inherits all of the above properties of the resource, i.e. has an identifier.

Each information object in the DL consists of the following objects:

- Metadata - is an object whose main purpose is to provide information about the resource;
- Abstract - is an object whose main purpose is to annotate the resource or a part of it. Examples of such abstracts include notes, structured comments and links. The objects of abstract helps to interpret the resource and contains detailed explanations, or information about how to use the resource.
- Information content - is an object that can be absent and can be used independently as a primary resource.

Information objects can also be complex objects and can be grouped into collections of information objects. Collections, in turn, are also information objects, they inherit all aspects of modeling information objects and means of their maintenance, for example, they can be annotated.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Information objects describe the basic nature of the scientific information space, such as publication, document, person, vocabulary, function and user, as well as the links between them.

IV. THE MAIN ENTITIES OF DL

Basic entities. The core or core of the information system model is constitute basic entities with which the DL operates. The basic entities are: Document, User, Function and Person. Each of the base entities has its own set of attributes. Instances of the underlying entities can be associated with named relations between themselves both within one entity and with all other entities.

The document - the main "object" with which the digital library operates - is a complete information object that represents a structured description of the real entity (object, subject, fact or concept), the totality of which constitute the information content of the system. The document has some standard set of attributes, described by metadata, and functions, depending on the class and type of document, and allows for unambiguous identification.

We will distinguish between the following classes of documents: publication, dictionary entry, key term. Each class of documents can have many types and subtypes. For example, the class of "publication" documents can have types: monograph, article, preprint, etc.

The collection - a set of documents of a certain class, united by a semantic feature and having the same structure of metadata. For example, many key terms that determine the composition of logical semantic categories (facets) covering a selected subject area of interest to the user may be a collection. Collections will also be a controlled dictionary, composed of descriptions of organizations.

The user - contains all objects that are external to and interact with DL: people and inanimate objects (programs or physical instruments or even other DL can be among DL users). For rights and functionality, users are divided into administrators and end users.

The function - is a specific processing task that can be implemented on a resource set or a single resource as a result of the actions of an individual user. The function represents the most voluminous and most open part of the model, since it covers the entire processing of resources, as well as the actions of users in the DL. In this model, the functions define four activities: Resource access, Resource management, DL management, DL configuration.

The persons - actors / individuals (both living and dead). Persons have an identifier and have metadata. Some people can have a property, be authors of publications, be the author of the fact described in the dictionary entry or simply have something to do with the concept (for example, as Plato to the concept of ontology). Persons are also linked to the thematic thesaurus twice: the keywords found in the person's publications by the indexing algorithm; keywords, characterizing the person, put by the expert.

The organization - a group of people engaged in certain activities to achieve common goals. In the model, the organization will be represented in the form of research institutes, educational institutions, publishing houses, etc.

V. THE METADATA

The description of the entities in this model is represented by metadata. Metadata - structured information that describes, explains and indicates the location of the information resource [7]. Metadata is needed to solve the following tasks:

- providing information about the object, its content, structure, methods of use, etc.;
- collection and systematization of information about objects, classification of objects;
- a choice from a set of objects of a certain subset according to formal characteristics and a comparison of objects according to formal characteristics;
- intrasystem technological tasks associated with ensuring the preparation of facilities, placing objects, etc.;
- external technological tasks, connected, first of all, with the exchange of data with external information systems.

The main types of metadata are:

- Descriptive metadata - are metadata that describe the content and properties of a resource, for example, bibliographic data, abstract, resource identifiers, whose main task is an unambiguous representation of a digital object for the outside world and in various applications.
- Structured metadata - are metadata that characterize the overall structure of a resource and its components, volume and other resource properties;
- System or Administrative metadata - serve to provide information resource management and administration of information resources, for example, the date of creation or modification of the resource, the owner's identifier, etc.

The metadata schema - is a set of metadata elements, each of which has a certain name and semantics, takes values with the set semantics or values from a controlled dictionary. In accordance with the recommendations of Dublin Core, the information object must have a basic set of attributes [2]. The set of attributes of the object is expanded depending on its type.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

A particular type of metadata is metadata describing the relationships and relationships between resources - documents.

The relation - is a relationship between an instance of a certain entity and what is related to it. The number of types of relations in the information system is determined based on specific goals. In the real world, their number tends to infinity. From the point of view of the information needs of users, we will be interested in relations not only between the documents, for example, "Publication - Publication", "Publication - Dictionary article", "Publication - Keyword", and relations "Publication - Person", "Persona - Dictionary Article "and so on. Links exist between all classes of documents.

Depending on the conditions of use, the relationships between documents are divided into the following types: thesaurus relations, semantic relations and associative relations:

- Thesaurus relations: the relations used in the description of information retrieval thesauri are hierarchical relations and the relation of association. The main hierarchical relationship is the generic relationship (parent-child, wider-already, higher-lower, part-whole). The main purpose of establishing associative relations between documents is to indicate additional links. [8, 9] Thesaurus relations are specific to the relationship between key terms, are much less commonly used in assigning relations between publications and vocabulary articles. For the implementation of the thesaurus relations, the Zthes data schema [9] was chosen, as the most advanced of the standard schemes [10].
- Semantic relations: named relations between documents and any other class, for example, "The publication is devoted to the Fact described in the Dictionary article"; "Person is the author of Publication"; "The publication is dedicated to Persona".
- Association relations: the relationship between two documents that are close in content, for example, keywords in the description of the Publication and the Dictionary article.

VI. THE DOCUMENTS

A. The key terms

Key terms are not used independently, but as part of collections. Two types of collections of key terms are used: Thematic dictionaries - Thesauri and Controlled dictionaries - a special entity that is designed to fill certain attributes (metadata) of other documents (names of cities, countries, organization names, subject heading names). Key terms are linked between with thesaurus relations. One of the differences of this information system from other systems is the support of classifier dictionaries, which are the basis of the system.

The identification of key terms are made using an identifier that is unique within one collection and using qualifiers, calculated from the term name, its language, abstract and term name in its normal form.

Each key term is associated with a dictionary entry, which provides its detailed description and the delineation of polysemy.

B. The publications

Publication is the embodiment of the results of the intellectual realization of a work in the form of an alphanumeric record that has output data (a bibliographic description). The main purpose of the publication is the dissemination of information contained in it. [11].

The main types of publications: Book, Article, Normative document (for example, standard, legislative act), etc. In turn, the publication type can have a subtype (for example, the type of article has a subtype: in the collection, article in the magazine, article in the newspaper, etc.). The most complete list of publication types are given in GOST 7.19-2001 [12] and in [11].

The publication has a basic set of attributes, based on the Dublin Core data schema, extended in accordance with the requirements of MECOF [12]. The type of publication depends on the set of mandatory descriptive metadata and the rules for their display. Publication is the only class of documents that can have information content. Typically, the information content (full text) is an external object in relation to the DL stored in the digital repository [1, 2], and in the metadata of the system is represented by a reference to the resource.

Publications can be associated with the following classes: persons, vocabulary articles and key terms. A publication can have a named relationship with the person class: authorship and character.

In addition to links with various collections of key terms that make up controlled dictionaries, the publication always has a connection with the thematic collection (thesaurus), and three times: the keywords put by the authors; Key words found in the publication by the indexing algorithm; Key words, characterizing the publication, put by the expert.

C. The dictionary entry

A dictionary entry is a document that describes the key term, concept or fact.

Fact - the characteristic of the entity described in the ontology of the information system, represented in the text of the document, as a single value of the data. The fact can be

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

extracted from the information content of the object or determined by the expert. The fact can determine how the properties (attributes) of the object, and its relationship with other objects [2].

Dictionary entry also contain links to descriptions of individuals, publications, links to key terms. Relationships with persons are semantic or associative, the relation with publications is associative (for example, an additional description). Relations with key terms: keywords found in the description (text) of the dictionary entry by the indexing algorithm; Key words that characterize the dictionary article delivered by the expert.

VII. THE USERS

User - the person involved in the operation of the information system to obtain information or perform a specific task (adding a resource, searching for a resource, etc.).

The user is an information object (resource) and, consequently, inherits all of its properties, i.e. Each user is provided with a unique identifier and has metadata.

Also, the user contains all objects that are external to and interact with DL: people and inanimate objects (programs or physical tools or even other DL can be among DL users).

In this model, the user is divided into two subclasses: Administrator, End user. The following is a complete classification of users of the information system.

The administrator is responsible for the operation of the information system and performs work on the management functions of the DL. Also defines end users and their rights.

The end user accesses the information system for adding / receiving information and often forms the most numerous group.

The end user is divided into: Content Manager, Owner and Reader.

Content Manager is an information system user who can manage the resources of all Owners.

The owner is the user of the information system, which has the rights to manage (add, change, delete) resources. The owner can authorize the management of his resources to other Owners.

The reader is a kind of user who has access to the resources of the information system. The main activity of the reader is to search and view the resource.

Readers are divided into authorized and unauthorized readers, as the resources of the information system can be closed or open.

Authorized readers can view all the resources and collections that are defined by the information system administrator.

Unauthorized readers have access only to open resources and collections of the information system.

Each user has a profile (account). The profile contains the main information and information about the user, as well as the information necessary when authorizing the user: last name, first name, patronymic, access rights, email address, organization and login, password.

Users can also be members of certain groups that are created to simplify the management of access to the DL resource.

VIII. CONCLUSION

In this paper is described an information model of the digital library for supporting scientific and educational activities. The model is typical and gives the opportunity to fix the rules and formulate the requirements for the system, can be used to develop information systems.

It is given a brief review of models, which describe scientific information resources that more or less are linked with scientific and educational activities.

The main entities used in the model, as well as their classes and subclasses, types of metadata and relations are distinguished and considered in detail. The model is based on the concept of the document as the main essence of the scientific information space, which includes such entities as publication, person, organization, fact, key term, etc., as well as relationships (relations) between them. The advantage of this model is language independence, supporting of multilingual thesaurus and the possibility of using different classification schemes.

REFERENCES

[1] A. Fedotov, O. Fedotova, “A model of information system to support scientific and educational activities,” Computational and Informational Technologies in Science, Engineering and Education CIT 2013: Proceedings of the International Conference. Ust’-Kamenogorsk, vol. 2: Computing technology: East Kazakhstan State Technical University, pp. 249-265, 2013.

[2] A. Fedotov, O. Zhizhimov, O. Fedotova, V. Barakhnin, “A model of information system to support scientific and educational activities,” Vestnik of Novosibirsk State University, Series: Information Technologies, vol. 12, № 1, pp. 89-101, 2014.

[3] Yu. Shokin, A. Fedotov, O. Zhizhimov, O. Fedotova, “The control system of digital libraries in IRIS SB RAS,” Infrastructure scientific information resources and systems: Collection of scientific articles of

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

the Fourth All-Russian Symposium. E. B. Kudasheva, V. A. Serebryakov (eds.), Moscow, vol. 1, pp. 11-39, 2014.

[4] V. Reznichenko, G. Proskudina, K. Kudim, "Conceptual Model of Digital Library," Digital Libraries: Advanced Methods and Technologies, Digital Collections: Proceedings of the XI All-Russian Research Conference RCDL'2009, Petrozavodsk: KRC RAS, pp. 23-31, 2009.

[5] Reference Model for an Open Archival Information System (OAIS), Recommended Practice, Issue 2, CCSDS 650.0-M-2 (Magenta Book) June 2012.

[6] A. Bezdushnyi, A. Bezdushnyi, V. Serebryakov, V. Filippov, "Integration of metadata of the Unified Scientific Information Space of the Russian Academy of Sciences," Moscow, p. 238, 2006.

[7] M. Kogalovskii "The metadata, their properties, functions, classification and presentation tools," Proceedings of the 14th Scientific Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" - RCDL-2012 Pereslavl-Zaleski, 15-18 October 2012.

[8] V. Solovyev, B. Dobrov, V. Ivanov, N. Lukashevich, Ontologies and thesauruses: Book, Kazan, Moscow: Kazan State University, Moscow State University after M.V. Lomonosov 2006.

[9] ANSI/NISO. Z39.19: 2005 Guidelines for the construction, format and management of monolingual controlled vocabularies. NISO Press: Bethesda, MD, ISBN:1 880124 65 3. 2005.

[10] A. Fedotov, I. Idrisova, M. Sambetbayeva, O. Fedotova, "Using the thesaurus in the scientific and educational information system," Vestnik of Novosibirsk State University. Series: Information Technologies, vol. 13, № 2, pp. 86-102, 2015.

[11] Functional Requirements for Bibliographic Records, Final Report / IFLA Study Group on the Functional Requirements for Bibliographic Records. – München: K.G. Saur, 1998. (UBCIM Publications, New Series; v. 19). http://archive.ifla.org/VII/s13/frbr/frbr.htm.

[12] GOST 7.19-2001 Format for data interchange. Contents of record p. 58, 2002.

  ISSN (Print): 2204-0595
ISSN (Online): 2203-1731