Conceptual model of information system for resources storage *

M. A. Sambetbayeva[1], O. A. Fedotova[1,2], A. M. Fedotov[1,3], A. G. Batyrkhanov[4]

[1] Novosibirsk State University, Pirogova 1, 630090 Novosibirsk, Russia
[2] State Public Scientific-Technological Library of SB RAS Voshod 15, 630102 Novosibirsk, Russia
[3] Institute of Computational Technologies of SB RAS ak. Lavrentieva Ave., 6, 630090 Novosibirsk, Russia
[4] L. N. Gumilyov Eurasian National University Satpaeva 2, 010000 Kazakhstan, Astana
{Madina_jgtu,o4f8,batyr.khan78}@mail.ru
http://www.ict.nsc.ru/

Abstract One of the main current trends in the socio-economic development of society is the informatization of all areas of human activity. The tool for implementing this comprehensive process is information systems (IS) — systems for collecting, storing, processing, converting, transmitting and updating information. The article is devoted to the description of conceptual model of Information System. The model has to describe which entities can be represented at the IS, set the rules and relations (links) between entities. In particular it assumes the entity classification, abstraction, generalization. The purpose of Information Systems is to provide the end-user with necessary information. Depending on specific application area, the information systems may vary greatly in their functions, architecture and implementation. There is, however, one property, which is common to all IS. Any information system is intended for the collection, organization, storage and processing the information. The work is devoted to the problems of description and creation of information models of the IS intensively developing class — digital libraries (DL).

Keywords: information system, digital library, model, metadata, data, information, knowledge

Introduction

The article is devoted to the description of conceptual model of Information System (IS). The model has to describe which entities can be represented at the

* This research is supported by RFBR Grant of the Russian Federation (grant No. 18-07-01457) and Grant funding for scientific and scientific and technical research for 2018-2020 MES RK (№ AP 05133546).

IS, set the rules and relations (links) between entities. In particular it assumes the entity classification, abstraction, generalization. The purpose of Information Systems is to provide the end-user with necessary information.

It is obvious that automated or computerized information processing is possible only in case when it can be described with some algorithm, i.e. in terms of formal data model and some system that perceives it. As noted by A. A. Lyapunov, "the information is always relative, it depends on what kind of IS it is perceived" [1].

IS for the purpose of research support operate with publications, electronic documents, electronic collections, ontological descriptions, etc. The resources may be inaccessible due to problems with their search and identification. Semantic relations between information resources increase their value and provide additional opportunities for information retrieval and identification. The data integrated into the semantic space is a set of knowledge about a subject area as a semantic structure. This is the basis for structure investigations of scientific knowledge. One of the main tasks of integration is to set the links between certain scientific facts (for example, "what does the term cybernetics mean" or "who is the author of this article" and information system entities (persons, facts, data, documents, publications, key terms, etc.) [2].

Depending on specific application area, the information systems may vary greatly in their functions, architecture and implementation. There is, however, one property, which is common to all IS. Any information system is intended for the collection, organization, storage and processing the information. That is why, any IS, including digital libraries, is based on data storage and data access environment [3].

The environment has to provide the level of storage reliability and access efficiency, corresponding to the area of IS application.

Thus, IS are intended for:

- organization of information storage (organization of repositories, support of data storing systems);
- information management (addition, modernization, changing the data);
- data access management (monitoring of fulfilment of the data access rules and regulations), data identification;
- search for information;
- information retrieval and provision to the user (computer application) in the required format;
- visulization (provision) of the information according to the user's requirements.

Digital Library Reference Models (DLRM) are of great interest. Let us note that the Russian term "electronic library"does not quite accurately reflect the content of the definition, it might be more precise to call this class of IS as "digital library"(English equivalent).

The conceptual reference models of the digital library (for example, DLRM [4] or OAIS RM [5]) are based on world experience and include concepts such as content, functionality, user, policy, quality and architecture, which leads to a

general understanding of the essence of digital libraries [6]. The main aim of the development of the DL model is to describe the fundamental concepts, essential objects and relationships, the standard functional and structural blocks and processes that make up the universe of DL. The reference model is designed to develop narrower models with a specific architecture for subsequent implementation in the form of a software system.

Definition of IS

In the case of IS definition, there is no unanimity of views. In daily use, information systems are usually referred to as various software and hardware complexes allowing to work with data, structured with a particular formal model.

At the same time, there is still some confusion as to the concepts of "information system"(IS) and "IS architecture"; it is not at all harmless and in practice it often prevents us from clear understanding what the subject of development is in a particular project: IS, or only its AC (automation complex) or an automated system (AS) entirely. As an example, let us consider two different definitions from GOST's:

Information system: A complex consisting of processes, technical and software tools, devices and personnel, capable to meet the required demands [7].

Information system: A system designed for the collection, transmission, processing, storage and delivery of information to consumers and it includes such main components as software, information support, technical facilities and maintenance personnel [8].

It is worth recalling that in the 1950s and 1960s the information sciences in the USSR occupied a prominent place, although their development was hampered by the specifics of the social structure of society. The turn in the subsequent years to work with data and the blurring of the term "informatics"led to the fact that much of the accumulated information was rendered, if not lost, then unclaimed, and the culture of working with information was lost. Today, for most users, the consumption of information services is more important, than its providing technologies.

But the most serious problem is the crisis in the sphere of information representation. The destructive moment is the absence of common generally accepted definitions in the field of information technology when it comes to "information"processing. First of all, since the time of Kolmogorov and Claude Shannon, at the engineering level, there occurred a confusion of concepts, the unification of representations about information and data or signals that encode this information, and in fact the "information"was understood as data sets.

Until recently, when IS were relatively simple, the lack of a clear distinction between "data"and "information"was not of practical significance. But with the advent of complex IS, where functions are distributed between a person and a machine, and with the development of disciplines such as decision support and knowledge management, more precise definitions of basic concepts "data "information"and "knowledge"are needed.

Today, there are two definitions of IS (technological and engineering):

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

Information system is a software and hardware complex, including computational and communication equipment, software and linguistic support, information resources, as well as serving (system) personnel.

Part of the real world, which is modeled by IS, is called its subject domain. Since the domain model, supported by the information system, is materialized in the form of properly organized information objects, it is called the information model (see Figure 1). Information objects are characterized by metadata describing the real object, and may be supplied with annotations. Information objects may have information content.

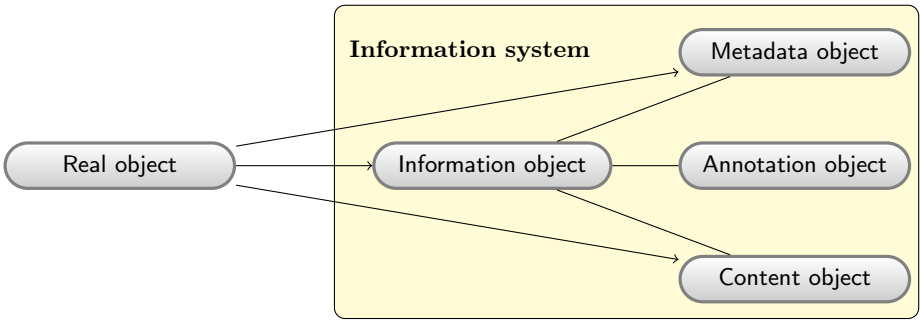

Figure 1. Information Model of IS

The above definition covers all IS classes, in particular factographic systems, which are based on database technologies and operate on structured data, and document systems that operate on documents in natural languages.

Information resource

An information resource is a concept that includes any essence of IS or DL. In the information space, all entities (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. An information resource is an abstract concept expressed by instances of one of its specializations. In particular, instances of the information resource concept are instances of an information object of any type.

So, basically, functionally, any IS presents functionally a system of managing information resources with its inherent functions (methods), relationships and links [2].

The implementation of an information resource (information object) is a unit of information, presenting a uniquely named set of data (see Figure 2) structured in the form of its inherent named attributes and methods that characterize its properties and relationships with other resources. Thus, each information resource must:

- have an identifier;
- be organized in accordance with the description of the resource (resources can be complex and structured, and in terms of their organization, they can be grouped into sets of resources that are treated as a single entity);
- be regulated by functions that manage its life cycle, characterized by a set of inherent attributes and methods describing its properties and connections with other resources;
- be expressed through an information object.

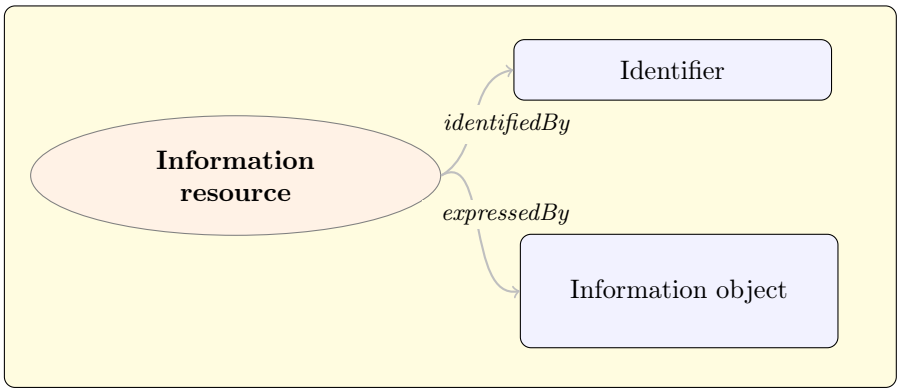

Figure 2. Determination of the Resource in IS

Each information resource is stored in some repository as a globally and uniquely named set of structured data (information about the resource, its properties, attributes, links) and, possibly, information content, for example, one or more presentation formats of the cataloged resource. These structured data describing the resource, and therefore called metadata, are used to get an idea of its properties, content, structure, search organization, ways of using, and so on.

Let us give the definitions.

In the IS, each information resource corresponds to the information object that is a traditional secondary information object containing a description of the primary resource, that is, an information object is an object that stores information about IS objects (physical objects, resources, information objects) (see Figure 3).

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

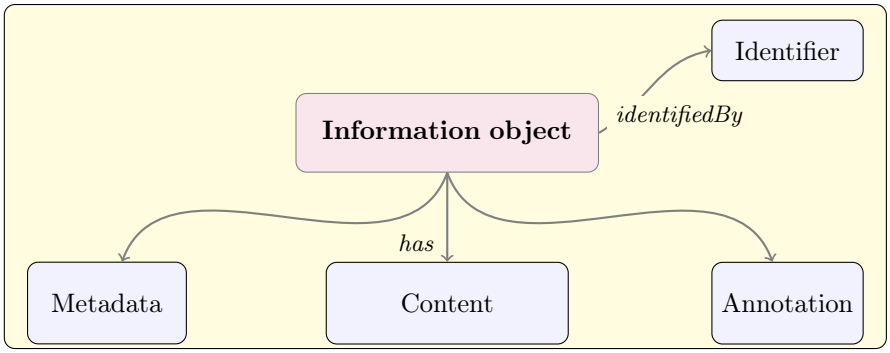

Figure 3. Information Object Model in IS

- *metadata* is an object providing information about the resource;
- *annotations* is an object annotating the resource or its part. Examples of such annotations include notes, structured comments and links. Annotation objects help to interpret the resource, contain detailed explanations, or information on how to use the resource.
- *information content* is an object that can be absent and can be used independently as a primary information object: for example, an image, a full text, etc. (primary resource) [2].

The information object is the most common concept in the system, representing an arbitrary unit of information in the IS. Information objects can also be complex objects and can be grouped into collections of information objects, which, in turn, are also information objects. Collections inherit all aspects of modeling information objects and the means of their maintenance, for example, they can be annotated.

Information objects describe all the classes of essences of the scientific information space, such as publication, person, key term or concept, dictionary entry, fact, function, organization, user, etc., and the links between them [10],[11].

Definition of the document

A document is a complete information object, presented in digital form, supplied with structured metadata, having some standard set of attributes and functions and allowing unambiguous identification.

Accordingly, the definition of the Digital Library (DL) can be given as follows: A set of structured cataloged collections of disparate electronic (digital) documents, equipped with navigation and search tools.

DL is able not only to provide a multilateral search in the catalog, but also to provide the user with a directly found resource (publication, document, photo, fact description, etc.), as well as additional information resourse, for example, authors, bibliography, organization, etc. And a great interest in the systems

of this class is explained by the actual needs of society and the availability of developing opportunities to meet them. In this regard, it is possible to formulate the main goals facing digital libraries:

- providing access to information;
- preservation of scientific and cultural heritage;
- increase the effectiveness of scientific research and training.

In most cases, the DL is a website where various texts (literary, scientific and technical, including publications, computer programs, digital maps, etc.) and media files are accumulated. The difference between DL and the magazine website is that the DL is not divided into numbers (issues) and is updated as new materials appear. The difference between DL and the site of free publications is that DL is usually selected by the project administrator according to certain rules and does not always provide a communicative environment around the published texts.

In existing developments of digital libraries, as a rule, the search and access to information are provided only through visual graphical user interfaces. This is good for a human user, but very bad for a user-system. To provide search functions outside graphical interfaces, special network services and query languages are required. Ideally, all information systems should support a single search profile and a single query language.

A digital library can be called a system of network services that provide access to digital content, united by a single system for managing this access [11]. In addition, some organizations that take responsibility not only for the implementation of the functions of managing digital content, but also for providing access to it for all interested parties. This definition of the digital library fully corresponds to the definition of the traditional library as an organization in the system, for example, of the Ministry of Culture [12].

Let us single out three concepts for distinguishing what is usually understood by the term "digital library"(see Figure 2):

- Digital Library (DL) is a specific DL with its users, policy, content and responsible organization, which can be virtual.
- DL system (DLS) is a software system (for example: DSpace [13]), which is based on a certain (possibly distributed) architecture and provides all the functionality required by a particular digital library. Users interact with the digital library through the appropriate digital library system interfaces.
- DL Management System (DLMS) is a generalized system software (for example: DLMS ICT SB RAS [2]), that provides the appropriate software infrastructure (I) for administering a digital library system, including a set of functionality that is considered fundamental for digital libraries and (II) allowing to integrate additional software offering specialized or enhanced functionality for creating and managing DL.

In DL each resource is defined as follows (see Figure 2):

– has an identifier;
– organized in accordance with the format of the resource, the format is described here by structural metadata that are an ontology (the resource can be complex and structured, since it, in turn, can consist of less resources and have links to other resources);
– described by structured metadata and annotations;
– can be characterized by quality parameters;
– can be regulated by policies, the policy of managing its life cycle;
– expressed through an information object;
– can be described or supplemented by an information object.

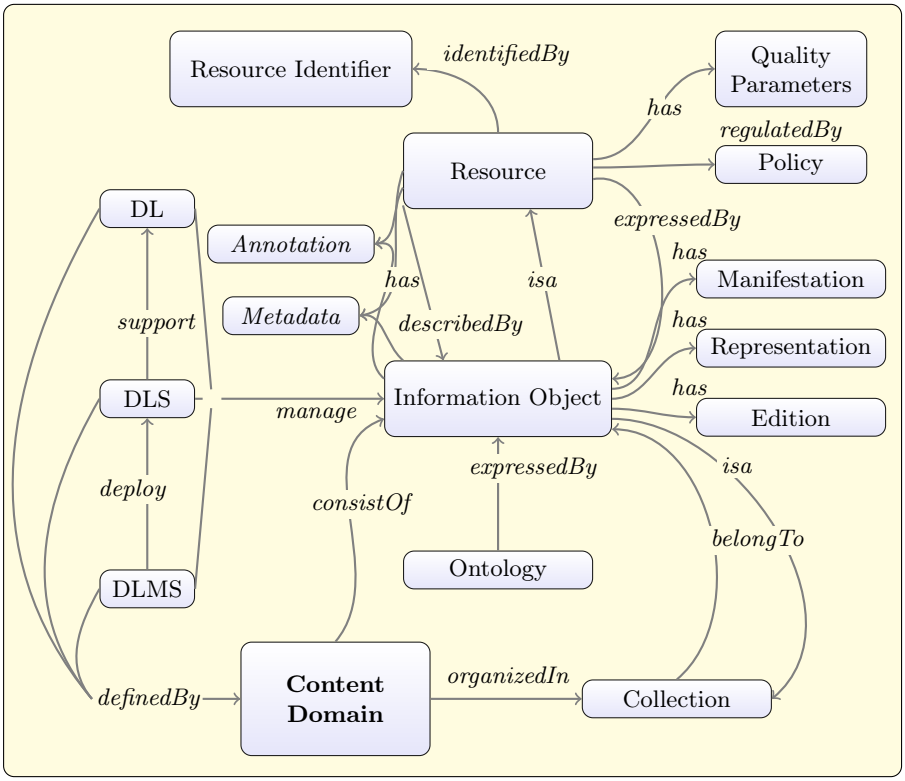

Figure 4. The content domain of DL

From an organizational point of view, resources can be grouped into collections of resources that are treated as a single entity.

Information model of DL

Based on the goals of DL and the analysis of existing IS aimed at supporting scientific research, the following functional requirements for the model of a scientific digital library can be formulated as follows [14]:

- Reliable long-term and non-stop storage of information.
- The relevance, completeness, reliability of origin of documents.
- Historicity of information.
- Geo-referencing of information.
- Availability of a large number of dictionaries-classifiers (directories), to ensure the identification and classification of resources.
- Support for heterogeneous and poorly structured information resources.
- Support of interrelationships of information resources. Identification of information resources.
- Providing information to the user in the form chosen by the user.
- Availability of intelligent services to service user requests.
- Presence of program interfaces for support of analytical work of the user with the help of software applications.
- Support for interoperability requirements at both program and semantic levels.
- Support of work with external sources (for example, catalogs of libraries and magazines, digital depositories of information resources, etc.). The most important conclusion from the foregoing is that the information model of the DL should be multilevel and consist of at least the following components [10]:
- data store — repository,
- a metadata server,
- application server (dispatcher),
- dictionaries — directories.

The repository is an independent system of long-term, reliable storage and access to heterogeneous digital objects, which is intended to provide electronic (digital) versions of documents (books, scientific articles, reprints, letters, images and other materials presented electronically), and providing a clearly specified method of management (data schema, model of operations), which includes, to some extend, the ways of accessing, sampling and manipulating information resources.

A collection is a collection of information objects (information resources), united by common properties (for example, belonging to the same class of objects, the same structure, the general thematic focus, etc.).

The metadata server must ensure the work with metadata — the cataloging of all information resources in accordance with generally accepted international standards. The application server must provide the basic services of DL. It provides the services necessary to generate information resources using and not using interactive user interfaces. Services allow you to use the metadata of other information systems in dialog mode and batch modes. Their functionality should

provide search and retrieval of metadata from other systems, converting received metadata into schemes and structures of the local system.

Directories — managed dictionaries (key signs, key terms) is a special kind of metadata reflecting the most essential properties of an information object and having the most importance from the DL's point of view. Specificity of words is determined by the terminology of the specific subject area, to which the DL is devoted. It is necessary to consider different types of key terms (key terms in the standard understanding, key terms describing the person, key terms describing the organization, key terms describing the time periods, key terms describing geographical concepts). This is a set of databases (normative dictionaries) containing information about authors and other persons (authoritative records), geographical points, cities, publishers relevant to a particular topic or section of DL (for example, to a scientific school), thematic classifiers, thesauri, rubricators, descriptions of the subject area and classifiers of documents.

The basis of the content of information security is information objects that represent the main types of entities:

- subjects: actor, person, organization, application, etc.;
- objects: publication, journal, document, fact, scientific result, event, project, photography, etc.;
- relationships: concept, key term, event, time, place, etc.

As a rule, the DL storage systems provide storing documents that are objects of works — the main type of information content objects (content), as well as some other objects related to them:

- descriptions of organizations, departments of organizations and publishers, where objects were created or published;
- description of people (in the scheme it is the person's essence), working in these organizations (departments) — authors of objects-works;
- a description of projects within which objects are created;
- scientific journals (periodicals) and conferences that publish them. A special type of object is an object collection can be applied to any aggregate (grouping, aggregation) of information objects. Information objects here can be of any type, i.e. collections can be both sets of subjects, and objects, for example, a set of organizations, journals, etc. The criteria for such collections can be determined, for example, by the generality of the location, the generality of the authors, the chronology, the theme, the origin or the membership, etc. Collections can contain any number of objects and the criteria for selecting these objects may change with time.

Metadata

The key moment in working with documents (information objects) is the use of metadata. Metadata is structured information that describes, explains and locates an information resource [15]. Metadata is needed to solve the following tasks:

– provision of information about documents, to get an idea of their properties, content, structure, ways of using, etc.;
– systematization of information about documents and system catalog maintenance;
– selection from a document set of a certain subset by formal characteristics and document comparison on formal grounds;
– intrasystem technological tasks related to supporting the document preparation, placing documents in the information environment, etc.;
– external technological tasks, first of all, related to data exchange with external information systems.

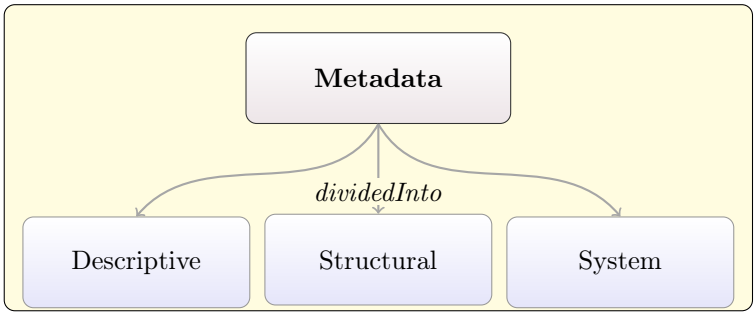

Figure 5. Basic Types of Metadata

Metadata reflects the most significant properties of the object, which have the greatest significance from the point of view of information system. Metadata is divided into three types: descriptive, structural and system (see Figure 5).

Descriptive metadata are metadata that describe the content and properties of an information resource, for example, bibliographic data the main task of which 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 the information resource and its components, volume and other properties of the information resource.

System or administrative metadata serve to provide a system for managing information resources and administering information resources, for example, the date of creation or modification of the resource, the owner's identifier, and so on.

The composition of attribute names, the restrictions imposed on their values, the set of rules that determine the structuring of attributes, their semantics are specified by a metadata schema. The rules for metadata streucturing in a certain sense are similar to the rules suggested by the ontology for constructing relationships between concepts. The rules that define the representation of

metadata in the information system, as well as the rules for their interpretation, form a metadata format.

A 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 managed dictionary, called an encoding scheme. In accordance with the recommendations of Dublin Core (DCMI) [16], the information object must have a basic set of attributes. The set of attributes of the object is expanded depending on its class.

The encoding scheme is the recording system or the rules for analyzing the values of the elements of the metadata. The value defined by the encoding scheme is a code (symbol) selected from a controlled dictionary (for example, a classification system index or a value from a set of subject headings) or a string of a specific structure (for example, "2000-01- 01 "as the standard date designation).

Controlled Vocabulary is a list of predefined codes, terms, words, phrases or notations designed to indicate subject headings or the composition of valid metadata element attribute values. All codes (terms) in the dictionary must have a unique definition.

A particular kind of metadata is metadata that unambiguously characterize (identify) documents that are necessary for the systematization of documents and for effective search, called authoritative.

Authoritative control (authority controls) is the provision of access to documents through a special class of metadata elements (proper nouns, subject headings, classification indices, activities, geographical names, names of document creators, etc.). The values of the elements of these metadata (which are keywords) are selected according to the encoding schemes from the managed dictionary.

There are problems in using authoritative data: first, authoritative monitoring must ensure the use of repeated geographical names and the names of document creators that are written identically, but denote different places or different people, and secondly, names can vary in time and space: for example, Germanized or Anglicized names and denominations are not authentic (authentic), so to study history you must also know the real names.

A particular class of metadata is the metadata describing the relationships and links between information resources — documents.

A link is a relationship between an instance of a certain entity and what it has been coordinated with. According to Aristotle, "there is what it is,"only "in connection with another, or in some other relation to another."The number of types of relationships in the information system is determined by 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 only in relationships between documents, for example, "Publication - Publication "Publication - Person "Publication - Vocabulary "Publication - Keyword "Persona - Vocabulary"and so on. Links exist between all classes of documents.

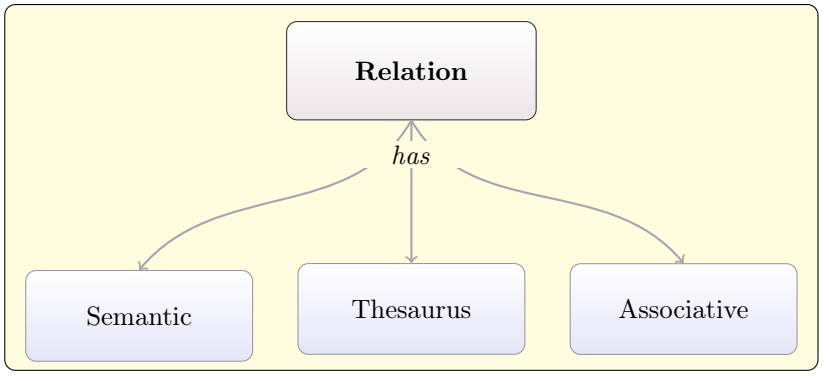

Figure 6. Types of Links

Depending on the conditions of use, the relationships between documents are subdivided into the following types: thesaurus relationships, semantic relationships and associative relationships (see Figure 6):

- thesaurus links: the relationships used in the description of information-subject thesauri are hierarchical relationships and the relationship of association. The basic hierarchical relationship is the subsumption relationship (parent-child, wider-narrower, higher-lower, part-entire). The main purpose of establishing associative relationships between documents is to indicate additional links [17]. Thesaurus relationships are specific for the relationship between key terms. They are much less frequently used when assigning relationships between publications and dictionary entries.
- Semantic links: named relationships between documents, for example, "Person is the author of Publication"; "The Publication is dedicated to Person"; "The Publication is devoted to the Fact described in the Dictionary article."
- Association links: relationships between two documents that are close in content, for example, keywords in the description of Publications, Persons, Dictionary articles.

In the information system, there are two possible ways of realizing the links (relationships) between the documents: hard and soft. Hard links are realized using DBMS by means of references to primary keys of record. Unfortunately, this type of communication is not protected against integrity violation (in case of incorrect modification or deletion of the record). Soft links are implemented through the matching procedure. This way of establishing links is protected from any violations of the integrity of the database and is quite convenient for users, since visual mnemonic definitions are used to indicate the need for communication.

Conclusion

As a result of the analysis of user needs, existing models and technological solutions, the main (basic) entities used in the model, as well as their classes and subclasses, types of metadata and relationships are specified and discussed in detail. The advantages of this model are the support of multilingual thesaurus and the ability of using different classification schemas.

References

1. Lyapunov A.A. On the relation of concepts matter, energy and information // In book: Problems of theoretical and applied cybernetics. — Moscow: Nauka, 1980. — P. 320–323 (in Russian).
2. Shokin Yu.I., Fedotov A.M., Zhizhimov O.L., Fedotova O.A. The evolution of information systems: from websites to information resource management systems // Vestnik NSU: Information Technologies, 2011. — Vol. 13. — N 1. — P. 117-134 (in Russian).
3. Fedotov A.M., Shokin Yu.I. Electronic Library of the Siberian Branch of the Russian Academy of Sciences // Information Society. — 2000. — N 2. — P. 22–31 (in Russian).
4. The Digital Library Reference Model / Candela L., Athanasopoulos G., Castelli D., Raheb K.El., Innocenti P., Ioannidis Y., Katifori A., Nika A., Vullo G., Ross S., Thanos C. // D3.2b DL.org Project Deliverable, April 2011 (in English).
5. ISO 14721:2012 Reference Model for an Open Archival Information System (RM OAIS) // Recommended Practice, CCSDS 650.0-M-2 (Magenta Book) Issue 2, June 2012 (in English).
6. Reznichenko V.A., Proskudina G.Yu., Kudim K.A. Conceptual Model of Digital Library [Electronic text] // Proceedings of the 11th All-Russian Scientific Conference RCDL, Russia, Petrozavodsk, September 17-21, 2009. — P. 23–31 (2009) (in Russian).
7. GOST R ISO/IEC 12207-99. Information technology. The life cycle processes of software. — Moscow, 1999 (in Russian).
8. GOST 34.003-90. Information technology. Set of standards for automated systems. Terms and definitions. – Moscow, 1990 (in Russian).
9. Sovetov B.Ya. Information technology: a textbook for high schools. — Moscow: Vysshaya Shkola, 2005 (in Russian).
10. Fedotov A.M. Methodologies of the distributed systems buildup // Computational Technologies: Special Issue. — 2006. — Vol. 11. — P. 3-16 (in Russian).
11. Shokin Yu.I., Fedotov A.M., Zhizhimov O.L., Guskov A.E., Stolyarov S.V. Digital Libraries — the path of integration of information resources of the SBRAS // Vestnik KazSU: Special Issue. — Almaty, Kazakhstan, Al-Farabi Kazakh National University, 2005. —N 2. — P. 115-127 (in Russian).
12. Zhizhimov Î.L., Mazov N.À., Fedotov À.Ì. Some notes on the evolution of digital repositories of traditional libraries for a full-featured digital libraries // Vestnik VSUES. — 2010. — Vol. 7. —N 3. — P. 55-63 (in Russian).
13. DSpace: an open source solution for accessing, managing and preserving scholarly works // MIT Libraries, HP Labs. 2007 (in English). Available at: `http://www.dspace.org/`
14. Fedotov, A.M., Barakhnin, V.B., Zhizhimov, O.L., Fedotova, O.A. A model of information system to support scientific and educational activities // Vestnik NSU: Information technology. — 2014. — Vol. 12. — N. 1. — P. 89-101 (in Russian).

15. National Information Standards Organization: Understanding Metadata // NISO Press, Bethesda, MD, 2004 (in English).
16. DCMI — Dublin Core Metadata Initiative (in English). Available at: `http://www.dublincore.org`
17. Lukashevich N.V. Thesauri in problems of information retrieval. Moscow: MSU, 2011. — 512 p. (in Russian).