Requirements For Information Systems To Support Scientific And Educational Activities[1]

Zhanna B. Sadirmekova[1], Oleg L. Zhizhimov [3], D.A. Tusupov[2], Madina .A. Sambetbaeva [1,2]
[1] L.N. Gumilyov Eurasian National University, Nur-Sultan, Kazakhstan
[2] Institute of Information and Computing Technologies of the Cabinet of Ministers of the Ministry of Education and Science of the Republic of Kazakhstan, Almaty, Kazakhstan
[3]Institute of computational technologies SB RAS

Abstract. The peculiarity of the development of modern society is characterized by an increasing volume and rapid obsolescence of scientific information. To increase the effectiveness of scientific research, scientists need access to information on the results of research carried out in the field of interest. Therefore, any scientific research usually begins with a search for scientific information about research in this field, but the search for the necessary information in an ever-increasing volume of articles, books, monographs, reports, patents is becoming more difficult. Scientists have to spend a lot of time searching and processing information that allows them to quickly get acquainted with the results of other studies and eliminate their duplication. The article gives a definition of information systems designed to support scientific and educational activities in terms of scientific communication. The task, subject area, subjects, objects, the basic functionality of the information system are determined, a list of the main types of information resources is provided. The paper analyzes the functional requirements for such systems.

Keywords: information system, scientific and educational activities, Z39.50, LDAP, DSpace, SRW / SRU.

Introduction.

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

The interoperability of any information system is understood as the degree of its ability to interact with other information systems, including with a person [1]. Ensuring system interoperability is impossible without strict compliance with relevant international standards and recommendations. In this case, the standards must meet:

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

In this paper, we consider a technological approach to creating a standard model of an information system designed to support scientific and educational activities. The developed model of the information system for working with materials related to 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.

In the process of scientific and educational activities, a lot of time and effort is spent working with literary sources, various materials and documents: search for the necessary documents, systematization and classification of documents in accordance with the task. To meet the information needs of modern users in the field of scientific and educational activities, it is necessary to support subtle functions of searching and classifying information, as well as viewing resources by categories (headings) and dictionaries-classifiers. The most important task is the task of systematization of resources (thematic classification), for which it is necessary to clearly define the composition of logical and semantic categories (facets) and key terms (concepts) that cover the selected rather narrow subject area of interest to the user.

The study formulated the main requirements for the information support system in scientific and educational activities. In particular, such a system should support reliable, long-term and protected from extinction, storage of information; include a large number of dictionaries-classifiers to ensure identification and classification of resources;

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

CEUR-WS.org/Vol-2569/short9.pdf

support poorly structured information resources, relationships between information resources; include intelligent services for servicing user requests, as well as software interfaces to support the user's analytical work; meet interoperability requirements, etc.

A set of the most General functional requirements for IS support of scientific and educational activities was identified [2]**.**

1) *A collection of information resources.* To collect information, you must use different data entry options:

– user input;
– collecting data on the Internet through special software agents (spiders);
– sharing data with others IS.

2) *The relevance of documents.* When information is automatically collected on the Internet, it may also accumulate irrelevant or poorly relevant information for this type of support for scientific and educational activities. The problem can be solved in the following ways:

– Creating detailed formats for presenting metadata about resources and structured reference books for thematic classification of resources. The research and education support system should embed descriptions in metadata on web pages and provide interactive tools for users to create metadata in a specific format when placing resources.
– Separation of information resources depending on the option of entering the system (posted by experts / users and a spider, as well as indicating the degree of reliability of information based on its source;
– Indication of search space information search and classification of information, as well as setting criteria for assessing the quality of the entered information;
– Using resource classification schemes according to user needs and resource classification according to these schemes.

3) *Relevance, completeness, and reliability of the origin of documents*. Ways to address issues of relevance and completeness are similar to ways to address the problem of resource coverage. Methods for determining the authenticity of information origin are as follows:

– for interactive input, information is entered only by authenticated users;
– for automated collection systems, placing restrictions on the scope of the agent that collects information;
– to exchange information with other IS details of the job filters on the imported information resources;
– for all input methods, all entered information must be checked and classified.

4) *Use of intelligent services for processing user requests.* User request processing services must provide attribute search, full-text search, resource browsing by category, and semantic search (optional).

5) *Knowledge extraction.* Using partial automation of knowledge extraction. The approach is based on the representation of the meaning of the text in the form of a semantic network, the principle of construction of which is based on the use of the frequency of joint occurrence of concepts in the text. The network is presented to the user as a thematic tree (a tree of key terms and related concepts), which allows navigation and significantly facilitates the process of researching the text and searching for the required information. This approach is also used for solving tasks such as automatic abstraction, thematic classification and clustering of texts, semantic search, etc.in Addition, the following requirements are imposed on the is of support for scientific and educational activities that work with different types of information resources.

6) *Support for non-centralized information system architectures.* This requirement is a prerequisite for the completeness, authenticity and relevance of the information. The experience of using IP to support scientific and educational activities has shown the complexity of creating centralized scientific systems that cover scientific information in a particular field of science, or in a country.

7) *Structuring of the information space.* To support complex information search and classification functions, in addition to storing a full-text description, you must implement attribute search, full-text search, and resource browsing by category and classifier dictionaries. The choice of classifiers is determined by the degree of specialization of the system.

8) *Adaptive presentation of information.* In order to improve the search speed and accuracy of information selection by the user without losing the quality of search, the research and educational support IS must take into account the users ' requests, their competence when working with the research and educational support IS, and time limits. The support system for scientific and educational activities should allow the user to obtain various levels of abstraction when presenting information, from short descriptions for maximum quick search, to very detailed descriptions of information objects.

9) *The historicity of the information.* The specificity of scientific information is its rapid obsolescence and loss of relevance. For many types of information resources, it is important to store all information about all changes and be able to restore the state of the resource at any time. For example, information about authors may change over time when a person changes their last name or place of work. It is necessary to take into account the re-formation and renaming of organizations, names of geographical objects, which may also change. Therefore, it is necessary to take into account the time factor and use up-to-date information for entities associated with time intervals. When recognizing entities, it is necessary to ensure that queries are executed at a certain point in time in the past, that is, creating a slice of the truth of information about entities on an arbitrary date.

10) *Archive.* As noted above, most scientific information is rapidly becoming obsolete. But there are information resources that need to be accessed for a long time. These include, for example, documents that have long-term legal force, patents, or multimedia information about historical events that may be in demand at any time. In addition, scientific reports of institutions, speeches of scientists can also have a huge historical value, acquiring its significance over time. Therefore, systems must support long-term storage of information resources with the ability to restore them.

When working in a distributed environment, the requirements for supporting research and

educational activities are met:

- support for accepted 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.

The main subsystems of the information system and approaches to their implementation are considered. Let's consider the technology of building a prototype of an information system for supporting scientific activities in accordance with the above requirements.

The main functions of information systems include the functions of collecting and registering information resources, saving them, processing, updating, and processing user requests [3]. **Collection and registration of information resources.** In this information system, data entry is performed semi-automatically and automatically, and information resources are paper or electronic publications.

When implementing these functions, you need to solve the tasks of cleaning, verifying, compressing data, and converting data from one format to another.

The operation of the environment is based on the use of the Z39.50 and LDAP protocols. At the same time, there are mechanisms for converting data from object schemas to the abstract schema of the Z39. 50 Protocol. The virtual environment consists of a registry of objects and resources, the main Z39. 50 server, several functional modules, and a web interface with public and administrative sections for accessing various features of the environment. For each source, a separate Z39.50 server is installed that converts data from the source schema to an abstract data schema.

Storage of information resources. The information model of the information system being developed should be multi-level and consist of at least two components [3, 5]: the data storage subsystem and the information resource management services subsystem. Moreover, the data storage function must be separate and independent of other functions and services of the system. The data storage subsystem is one of the most important components of the system and is intended only for providing the" function " of long-term storage of information resources. Services can change, and data must be stored securely and forever. The DSpace system was selected in the information system. The DSpace system has a number of attractive features:

1. Organize storage of digital objects (images, media files, documents in various formats, etc.). Version 5.0 of DSpace initially "understands" more than 70 file types, but this number can be easily expanded depending on the needs of a particular repository.
2. Provide these objects with metadata according to different data schemas. A specific model based on the Dublin Core schema is fixed for the underlying metadata organization (http://dublincore.org/documents/dcmi-terms/) and its extensions (http://purl.org/dc/terms/). it is Possible to use other additional schemes, including those defined locally. DSpace uses PostgreSQL or Oracle relational database to store metadata.
3. Создавать поисковые индексы, построенные в соответствии с заданными правилами на основе полей метаданных и контента цифровых объектов, например, текста, извлеченного из PDF документов. Для хранения и индексации цифровых объектов DSpace использует Apache SOLR.
4. Organize various hierarchical collections of digital objects. The structure and level of nesting of sections and collections can be any. A digital object can be registered in several collections at the same time.
5. Use WEB interfaces to navigate hierarchical collections, search, view metadata and digital objects, and manage and perform administrative functions. DSpace provides two groups of interfaces: JSP-based and XML-based.
6. Control access to the repository content and its functions (depositing, editing, etc.) with varying degrees of detail.
7. Maintain a list of repository users and user groups.
8. Perform user authentication using various technologies (local, LDAP-based, etc.). The system stores information about users, supports authorization, and differentiates access to repository content by groups, network addresses, and based on the LDAP Protocol. when creating an information system, it makes it possible to use an existing user authentication system (rather than developing your own) and easily differentiate public and service resources, while leaving free access to metadata.
9. Import and export metadata and digital objects in accordance with common protocols and formats. The most commonly used XML-oriented protocols are OAI-PMH and OAI-ORE. However, the server that supports these protocols is included in the basic DSpace configuration. Using these protocols, it is possible to organize the synchronization of data between different repositories. For the basic data organization, a specific data model (DIM – internal DSpace data format) is fixed, based on the Dublin Core schema and its extensions. At a certain voltage, you can use this diagram to display the main elements of all currently used data schemas. The system uses metadata filters, which are used to convert metadata from an internal schema to schemas that are suitable for exchanging metadata with external systems based on XSLT transformations, to convert and index metadata in a variety of formats (MODS, METS, QDC, MARC, etc.). The list of formats can be expanded by adding new ones, including native generation,

converters, for example, for the MEKOF format.

10. To organize a batch of input-output data.

11. To organize the cataloging of new objects by the method of borrowing from other repositories and databases. The list of available external sources can be expanded by changing the configuration and / or adding new modules.

12. Control the data entered in the selected fields during cataloging, using internal and external directories. The most interesting feature is the ability to work with authority data from the Library of Congress and ORCID using the corresponding modules. Each source type has its own module. Configuration of existing modules is performed via the configuration file. The list of modules for authoritative control can be expanded by creating additional modules.

13. To migrate to new versions without disrupting the overall structure of the repository.

To better meet local requirements, numerous changes have been made to the basic DSpace system (expanding data schemas, expanding the range of exchange formats, the ability to work with geographical information, authoritative control, etc.).

In the upgraded system used, access to repository data is possible not only through the DSpace WEB interfaces, but also through the OAI-PMH, SOLR, SRW/SRU, and Z39.50 protocols. At the same time, support for SRW/SRU and Z39. 50 is provided by the DSpace link with the ZooSPACE system [6,7].

In addition to the main functions of information systems that are visible to users, there are additional functions, some of which are assigned to the staff of the information system and the subject area:

- managing distributed information resources, such as database fragmentation, data replication, and copy synchronization;
- protection of the physical integrity of information resources and their restoration in case of destruction;
- ensuring information security in the system;
– metadata management;
- administration of information resources;
- ensuring that the system adapts to changes in requirements for it and to changes in the subject area.

References

1. 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, - P.44-61.
2. Barakhnin V. B., Leonova Yu. V., Fedotov a.m. On the question of formulating requirements for building information systems of scientific and organizational orientation / / Computing technologies. - 2006. - Vol. 11. - Special issue. - Pp. 52-58.
3. ANSI/NISO. Z39.19: 2005 Guidelines for the construction, format and management of monolingual controlled vocabularies. NISO Press: Bethesda, MD, 2005. ISBN:1 880124 65 3.
4. Fedotov A.M., Tusupov J.A., Sambetbayeva M.A., Fedotova O.A., Sagnayeva S.K., Bapanov A.A., Tazhibaeva S.Z. Classification model and morphological analysis in multilingual scientific and educational information systems // Journal of Theoretical and Applied Information Technology. - 2016. - Vol.86, issue 1, - P.96-111.
5. Shokin Yu. I., Fedotov a.m., Zhizhimov O. L., Fedotova O. A. Evolution of information systems: from Web sites to information resource management systems / / Vestn. Novosibirsk. state University. Series: Information technologies. 2015. Vol. 13, vol. 1. Pp. 117-134.
6. Doerr M., Iorizzo D. The Dream of a Global Knowledge Network - A New Approach //ACM J. on Computing and Cultural Heritage, June 2008. - Vol. 1, N. 1, Article 5.
7. Candela L., Castelli D., Fuhr N., Ioannidis Y., Klas C.- P., Pagano P., Ross S., Saidis C., Schek H.-J., Schuldt H., Springmann M. The Digital Library Reference Model.// Technology-enhanced Learning and Access to Cultural Heritage. April 2011.