Nurlan TEMIRBEKOV[1], Dossan BAIGEREYEV[2], Almas TEMIRBEKOV[3], Andrzej SMOLARZ[4]

Kazakhstan Engineering Technological University (1), D. Serikbayev East Kazakhstan State Technical University (2), Al-Farabi Kazakh National University (3), Lublin University of Technology (4)

doi:10.15199/48.2019.05.19

Architecture of the distributed information system of the Almaty Academgorodok

Abstract. *The paper describes the architecture of an integrated distributed information system, which allows to preserve the result of the intellectual activity of Kazakhstan Engineering Technological University and a number of research institutes, located in the Academgorodok of Almaty, Kazakhstan. A description of the implementation of information system subsystems is presented.*

Streszczenie. *Artykuł opisuje architekturę zintegrowanego rozproszonego systemu informatycznego, który pozwala zachować wyniki działalności Kazachskiego Uniwersytetu Techniczno-Technologicznego oraz szeregu instytutów badawczych, zlokalizowanych w Kampusie Almaty w Kazachstanie. Przedstawiono opis wdrożenia podsystemów systemu informacyjnego. (**Architektura rozproszonego systemu informacyjnego w Kampusie Almaty**).*

Keywords: information system, institutional repository, system architecture
Słowa kluczowe: system informacyjny, repozytorium instytucjonalne, architektura systemu.

Introduction

In the era of information technology information has a significant impact on the direction of development in the scientific, technical, economic, socio-cultural and other areas of life of any community, state or organization. Currently, information is one of the most significant resources, the preservation, rational use and development of which is one of the strategic directions.

There are several research institutes in the Almaty Academgorodok that have been conducting research in various areas of the agricultural industry for several decades: the Institute of Human and Animal Physiology, the Kazakh Research Institute of the Processing and Food Industry, the Kazakh Research Institute of Fruit and Viticulture, Kazakh Research Institute of Soil Science and Agrochemistry named after U. Uspanov, Kazakhstan Engineering Technological University, the Institute of Zoology, as well as Research Institute of Microbiology and Virology, Institute of General Genetics and Cytology, Institute of Seismology, and others. Significant amounts of information obtained as a result of research of these institutions, their continuous increase and heterogeneous nature of storage and distribution in many ways, the lack of unified access to them create significant problems of their effective use. These problems lead to the need to find new approaches and solutions to the problems of creating a repository of information resources, their organization, means and methods for users to access them. Today such approaches are called “digital” or “electronic” libraries [1, 2].

Recognizing the need to create a unified research and education cluster, the general decision of the leaderships of these research institutes and Kazakhstan Engineering Technology University set the goal to create an integrated distributed information system of the Almaty Academgorodok, which allows to keep the result of intellectual activity of the above research institutes in their current form and provide access to them through Web technologies.

One of the current areas is the creation and use of distributed computer systems for intensive work with data, which is important both for solving new scientific problems, using large amounts of scientific data generated by modern measuring tools, and economic and social problems based on big data and their technology of processing and analysis. One of the main results of scientific activity is the creation and accumulation of experience of previous generations.

The information system described in this article is focused on meeting the needs of the participants of the research and education cluster, is based on advanced information technologies and is implemented on the basis of free software and unique software modules.

Architecture of the integrated distributed information system

The software part of the distributed information system consists of the following subsystems presented in Fig. 1:
- Repository of digital objects;
- The subsystem for managing current research information;
- The subsystem of integration of distributed information resources;
- The subsystem of access to distributed information resources based on Web technologies.

The Digital Object Repository subsystem is intended for long-term storage of the results of scientific research institutes with the ability to search for information resources using metadata, full-text search, and statistics. The subsystem for managing current research information is designed to store information about research institutes, their employees, as well as information about their research activities (participation in funded projects, conferences, internships, etc.).The integration subsystem of distributed information resources is designed to import metadata from external sources (reference databases), as well as to provide metadata to internal resources (Akademgorodok portal) based on standard protocols. The subsystem of access to distributed information resources based on Web-technologies is designed to provide a standardized uniform user interface for all functions and modules included in the distributed information system. The portal provides remote access to information resources and services. The subsystem integrates the processes of providing access to information resources to all stakeholders of interaction. The following sections provide a detailed description of the components of an integrated distributed information system.

Implementation of the Digital Objects Repository subsystem

The following basic requirements were put forward to the software underlying the Digital Object Repository subsystem:

1. Ability to work with arbitrary documents, geographical maps, audio and video materials;

2. Flexible organization of resource storage (partitioning into collections, subcollections, etc.);

3. Flexible user rights: creating user groups; the ability to specify the access of users of a given group to a given set of objects according to the desired access method (download, view, edit, delete, change attributes); identification, authentication and authorization of users;

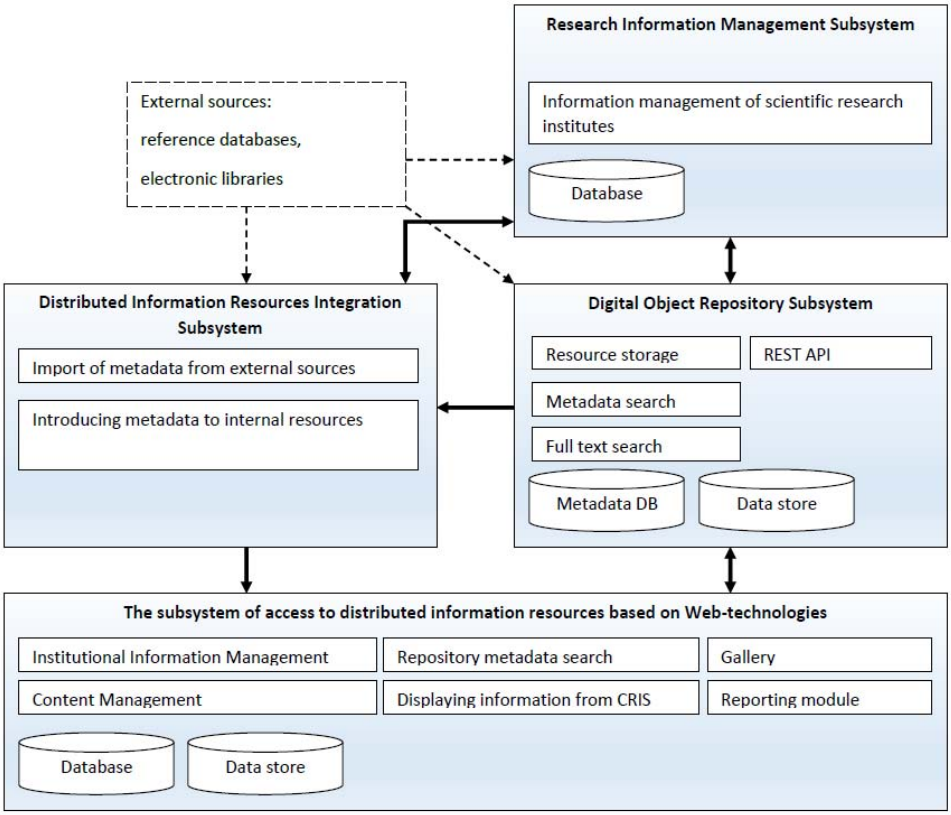

Fig.1. The relationship of the subsystems of the distributed information system

4. The ability to integrate distributed information resources based on standard protocols (Z39.50/SRU/SRW);

5. Availability of a software interface (API) for integration with internal resources;

6. Recognition of text in digitized materials for the organization of full-text search;

7. Collect statistics and provide various reports.

When choosing a digital object repository subsystem, the following institutional repositories and systems for creating electronic libraries were considered and the experience of their use in New Zealand, the Czech Republic, Sri Lanka and other countries were studied [2-9]: Ambra, Digital Commons, DSpace, ePrints, Evergreen ILS, Greenstone, Fedora Commons, Invenio, RODA, and VuFind. Analyzing the advantages and disadvantages of the listed systems, DSpace, ePrints and Greenstone systems were selected as the most satisfactory.

The strengths of Greenstone include the hierarchical structuring of each document, the automatic extraction of metadata from the document when it is loaded. However, this system supports only a limited number of formats: MS Word, Excel, RTF, HTML, Plain, PDF, ZIP, and MP3. Storage of geographical maps, as well as other results of scientific activities that have a more complex structure, in the opinion of the project executors, is not provided. The system provides wide opportunities for search queries: in addition to boolean operations and grouping words using parentheses, the search for words in the original form is supported.

As a result of the analysis, the DSpace repository of DSpace digital objects was selected. The strengths of DSpace include a more sophisticated system of user rights compared with the systems considered: various research institutes can have their own areas within the system. In each institute, certain officers responsible for pre-moderation can be appointed, i.e. users who have the ability to view and edit materials before they are included in the repository. DSpace, like the other systems considered, provides interfaces for integration with other subsystems based on open international standards. DSpace supports more than 70 formats of information resources. Materials in DSpace are indexed in Google Scholar. There is a large number of plug-ins to the DSpace system, expanding its capabilities, including the DSpace-CRIS research management system.

When building DSpace 6.2 software, changes were made to its configuration in order to adapt to the conditions established in the Republic of Kazakhstan. The standard DSpace metadata scheme based on the DCMI scheme is expanded by the following fields: “Journal in the list of Committee for the Control of Education and Science of the Ministry of Education and Science of the Republic of Kazakhstan”, the title, number, volume, issue of the journal, full bibliographic reference in accordance with State Standard, first and last pages of the article, etc. The list of resource types were supplemented with the following articles: “Article in the conference materials”, “Article in the newspaper”, “Report on Research & Development”, “Patent”, “Technical Report”, “Museum Object”, “Working Papers”. In addition, the ability to work with geographic information according to work was added. To support the process of filling full-text databases, the created metadata profiles are registered in the DSpace system, and the

workflows and the user interface of the system are configured in accordance with them.

The internal storage organization of resources in the DSpace system is structured as follows: seven communities are created in the repository, corresponding to research institutes. Each community, in turn, consists of several collections corresponding to the type of resource (articles, monographs, research reports, etc.)

Data entry into the subsystem is carried out:

- in an interactive mode through the built-in Web interfaces;
- borrowing data from other systems (DOI, PubMet, ArXiv, CiNii, CrossRef, etc.);
- in batch mode: import data in DIM, MEKOF, MARC21, DC formats, etc.;
- synchronization of data with external sources by OAI-PMH.

Access to the repository data is possible not only via DSpace Web interfaces, but also via OAI-PMH, SOLR, SRW / SRU, Z39.50 protocols. The latter is provided by DSpace communication with the ZooSPACE system [10].

Implementation of the Current Research Information Management Subsystem

As a research management system [11, 12], a free extension of the DSpace system, DSpace-CRIS was chosen. The system allows to store the following information:

- information on research organizations;
- information about the employee of research organizations, various spellings of his/her name, including different languages;
- links to profiles in various databases (Scopus, Researcher ID, ORCID);
- information on scientific activities (participation in funded projects, conferences, internships, etc.).

The system keeps statistics of publications for each scientist. The system is integrated with a repository of digital objects, which allows to view the publication of scientists. The CRIS-system allows to export information about the publications of the scientist in popular formats. The system allows for tracking changes on the page using RSS technology.

Implementation of the subsystem of Access to distributed information resources based on Web technologies

This subsystem of the distributed information system is designed to provide a standardized uniform user interface for all functions and modules included in the distributed information system. The portal provides remote access to information resources and services. The subsystem integrates the processes of providing access to information resources to all stakeholders of interaction. The subsystem involves:

- navigation through services;
- access to information about the project and regulatory legal and methodological materials;
- identification, authentication and authorization of users;
- providing basic information about research institutes, and their employees;
- providing information on the latest achievements of research institutes, upcoming events, and conferences;
- photo and video galleries;
- personal account, analysis of working time on the website, storing the history of requests;
- management of requests for the search for authors, the name of publications;
- management of requests for full-text resource search;
- export data in various formats;
- provision of various reports.

The interaction scheme of the Academgorodok portal with the rest of the components of the distributed information system is shown in Fig. 2.

The website operates on the Gunicorn WSGI HTTP server with the nginx HTTP server installed as a reverse proxy server.

Implementation of the Integrated Information Resources Integration Subsystem

As integrating software, the ZooSPACE distributed information system was chosen, which was developed by researchers at Institute of Computational Technology of Siberian Branch of Russian Academy of Science [13-15].

The ZooSPACE distributed information system integrates data from various information sources, providing access to heterogeneous distributed information in accordance with standard protocols (SRW/SRU, Z39.50). The system operates on the basis of original ZooPARK-ZS servers, LDAP servers and Apache WEB servers, providing end-to-end information retrieval in heterogeneous databases, extracting information in standard schemes and formats and displaying it [10].

The integration of the Academgorodok portal with the repository of digital objects is implemented using DSpace REST API, which provides a programmatic interface to communities, collections, elements' metadata and files.

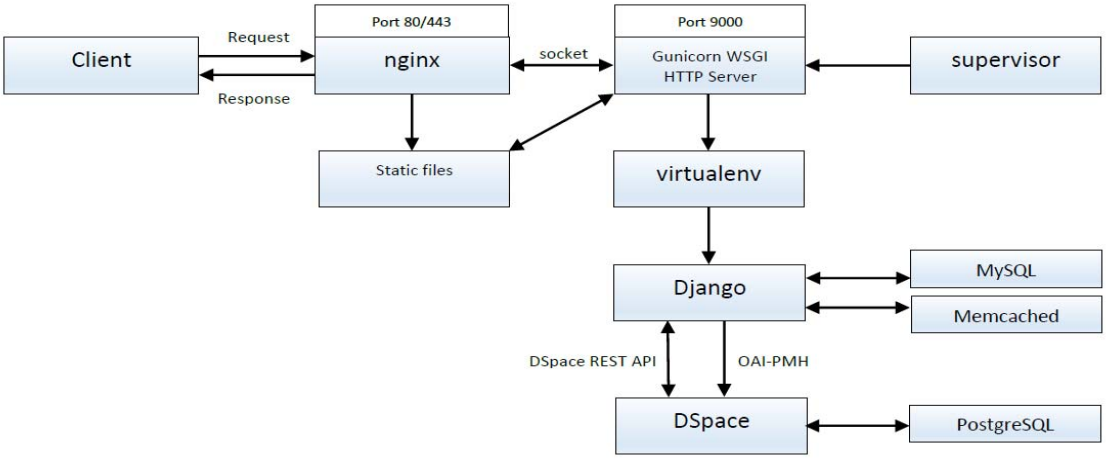

Fig.2. The scheme of interaction of the Academgorodok portal with the components of a distributed information system

The DSpace REST API is deployed as a standard Web application. An unprivileged user has been created in DSpace on whose behalf search queries are made. All requests are made using the curl utility; JSON is parsed in the Integrator application in the Django web application.

The results of the integration of the portal and repository can be traced on the following web pages:

1. Metadata and links to materials uploaded to the repository of digital objects are displayed on the scientist's profile page in the Publications in the digital repository section. To solve the problem of different spellings of the name of a scientist, an auxiliary table was created in the DBMS, which is currently filled by content managers in the administrative part of the portal.

2. Similarly, the list of all resources available in the repository of digital objects is displayed on the information page of research institutes. To ensure the interconnection of portal content managers, an identifier (handle) of the DSpace community associated with this research institute is entered.

3. Search of items by metadata is available.

The DSpace system has the ability to integrate with external databases ArXiV, PubMed, CiNii, ORCID, CrossRef and import metadata from the specified databases when filling the repository with new articles. Thus, content managers get rid of the need to manually fill in metadata.

Conclusion

Thus, the paper presents the architecture description of an integrated distributed information system of the Almaty Academgorodok. The results of the implementation of its subsystems are presented.

This developed system meets the needs of the participants of the scientific and educational cluster both in terms of informational content and in support of industry and language specifics, since it solves the main tasks imposed on these systems: ensuring a system of reliable long-term storage of digital (electronic) documents while preserving all semantic and functional characteristics source documents; providing a “transparent” search and user access to documents, both for review and for the analysis of the facts contained in them; organization of information collection on remote digital repositories that support the OAI-PMH, SRW/SRU, Z39.50 protocols.

The system fully provides the necessary computational resources for research and educational processes, simplifying the prospect of its further development, and allows to build an advanced IT infrastructure for managing intellectual capital, an electronic library, which will store all the books and scientific works of Kazakhstan Engineering Technological University and research institutes of the Almaty Academgorodok.

This research was supported by grant No. AP05131806, registration number 0118PK00411 of the Committee of Science of the Republic of Kazakhstan.

***Authors**: prof. dr. phys.-math. Nurlan Temirbekov, Kazakhstan Engineering Technological University, Al-Farabi ave. 93A, Almaty, Kazakhstan, e-mail: temirbekov@rambler.ru; PhD Dossan Baigereyev, Serikbayev str. 19, Ust-Kamenogorsk, Kazakhstan, e-mail: dbaigereyev@gmail.com; Almas Temirbekov, Al-Farabi Kazakh National University, Al-Farabi ave. 71, Almaty, Kazakhstan, e-mail: almas_tem@mail.ru; PhD., Prof. Andrzej Smolarz, Lublin University of Technology, Institute of Electronics and Information Technology, Nadbystrzycka 38A, 20-618 Lublin, Poland, e-mail: a.smolarz@pollub.pl.*

REFERENCES

[1] Saini O. P., The emergence of institutional repositories: a conceptual understanding of key issues through review of literature, *Library Philosophy and Practice*, 1774 (2018), 19 p.

[2] Franchke H., Gamalielsson J., Lundell B., Institutional repositories as infrastructures for long-term preservations, *Information Research*, 22 (2016), nr 757, 27 p.

[3] Hippenhammer C., Comparing institutional repository software: pampering metadata uploaders, *The Christian Librarian*, 59 (2016), nr 1, 6 p.

[4] Baughman K., McDowell. Institutional repositories in the Czech republic, *Gleeson Library Librarians Research*, 10 (2016), 29 p.

[5] Ravikumar M. N., Ramanan T., Comparison of greenstone digital library and DSpace: Experiences from digital library initiatives at eastern university, Sri Lanka, *Journal of University Librarians Association of Sri Lanka*, 18 (2014), nr 2, 76–90.

[6] Castagné M., Institutional repository software comparison: DSpace, ePrints, Digital Commons, Islandora and Hydra (Report), University of British Columbia (2013), 15 p.

[7] Cullen R., Chawner B., Institutional repositories in New Zealand: comparing institutional strategies for digital preservation and discovery, *Proceedings of the IATUL Conference*, 18 (2008), 11 p.

[8] Wancerz M., Wancerz P, History management of data – slowly changing dimensions, IAPGOŚ, 3 (2013), no 3, 55-56.

[9] Wojcik W., Kisala P., The application of inverse analysis in strain distribution recovery using the fibre bragg grating sensors, *Metrology and Measurement Systems*, 16 (2009), No.4, 649-660

[10] Shokin Yu. I., Zhizhimov O. L., Fedotov A. M., Information systems of ICT SB RAS, *Proceedings of the XVI All-Russian Conference DICR-2017* (2017). 11-18.

[11] Chudlarsky T., Dvorak J., A National CRIS Infrastructure as the Cornerstone of Transparency in the Research Domain. In: Jeffery, Keith G; Dvorak, Jan (eds.): E-Infrastructures for Research and Innovation: Linking Information Systems to Improve Scientific Knowledge Production, *Proceedings of the 11th International Conference on Current Research Information Systems (June 6-9, 2012, Prague, Czech Republic),* (2012), 9-17.

[12] Guskov A.E., Zhizhimov O.L., Kikhtenko V., Skachkov D.M., Kosyakov D., RuCRIS: A Pilot CERIF based System to Aggregate Heterogeneous Data of Russian Research Projects, *Procedia Computer Science*, 33 (2014), 63-167.

[13] Zhizhimov O. L., Fedotov A. M. Fedotova O. A., Building a typical model of an information system for working with documents on scientific heritage, *Bulletin of the NSU. Information Technology*, 10, (2012), nr 3, 5-14.

[14] Shokin Yu. I., Fedotov A. M., Zhizhimov O. L., Fedotova O. A., The control system of electronic libraries in IRIS SB RAS, *Infrastructure of scientific information resources and systems: Collection of scientific articles of the Fourth All-Russian Symposium*, 1 (2014), 11-39.

[15] Zhizhimov O. L., Fedotov A. M., Shokin Yu. I., Technological platform for mass integration of heterogeneous data, *Bulletin of the Novosibirsk State University. Series: Information Technology*, 11 (2013), nr 1, 24-41.