

On the Use of the Loud Platform in the Work of the Scientific and Educational Cluster

Nurlan M. Temirbekov, Tahir M. Takabayev, Dossan R. Baigereyev, Waldemar Wójcik, Konrad Gromaszek, Almas N. Temirbekov, and Bakytzhan B. Omirzhanova

Abstract—The process of designing and creating an integrated distributed information system for storing digitized works of scientists of research institutes of the Almaty academic city is analyzed. The requirements for the storage of digital objects are defined; a comparative analysis of the open source software used for these purposes is carried out. The system fully provides the necessary computing resources for ongoing 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 that is intended to store all books and scientific works of the Kazakhstan Engineering Technological University and research institutes of the Almaty academic city.

Keywords—IT infrastructure, cloud solutions, data center, server

I. INTRODUCTION

CURRENTLY, due to various historical circumstances, the IT infrastructure of many Kazakh scientific and educational institutions is an unsystematic set of heterogeneous and often disparate autonomous hardware and software systems. The disadvantages of this approach are obvious to IT specialists. For most of them, it is absolutely clear that its current state allows to maintain the educational process and ongoing research at a certain level, but at the same time, it is more of a deterrent than a basis for further successful development of domestic science and education in general. A similar situation exists in many developing countries that have focused on the development of digital technologies for personal computing resources.

At the same time, the global IT industry, after several generations of miniaturization and modernization of computer technology, has moved from the paradigm of personal computing systems to the idea of centralizing computing resources in data centers using client-server applications. The world experience has long been replete with successful examples of transferring the IT infrastructure of large commercial and educational structures to cloud principles using distributed data storage and processing technologies. This transition, while not a simple and expensive solution, at the same time allows to optimize the cost of maintaining the entire

This work was supported by grant funding of scientific and technical programs and projects by the Ministry of education and science of the Republic of Kazakhstan (grant no. AP05131806, 2018-2020).

First Nurlan M. Temirbekov and Tahir M. Takabaev are with Kazakhstan Engineering Technological University, Almaty, Kazakhstan (e-mail: temirbekov@rambler.ru, tt@ahost.kz).

Dossan R. Baigereyev is with S. Amanzholov East Kazakhstan State University, Ust-Kamenogorsk, Kazakhstan (e-mail: dbaigereyev@gmail.com).

infrastructure with the correct implementation, and most importantly, creates enormous prospects for owners. This is due to the fact that cloud solutions are more open to modernization and scale to any size. But, in the opinion of the authors of this work, the main advantage of cloud solutions for scientific and educational structures is their openness in testing any set of new test solutions. Classic cloud services IaaS, SaaS, and PaaS are intended primarily for limited time use in terms of marketing and business, that is, to attract third-party, fresh ideas and innovations in IT, literally, innovations on demand.

In this regard, to build a strategy for further development of the national IT infrastructure in terms of local adaptation of technologies and methodological experience, the analysis of the results of using a cloud platform in the construction and operation of a scientific and educational cluster based on the Kazakhstan Engineering Technological University (KazETU), the National Engineering Academy and several research institutes of the Almaty Academic city, located at a relatively small distance from each other in Almaty, may be useful. All works were performed with the support of grant funding of scientific and technical programs and projects by the Ministry of education and science of the Republic of Kazakhstan (grant no. AP05131806, 2018-2020). The relevance of the work is due to a number of state programs, including the state program “Digital Kazakhstan” for 2017-2020. Extensive research has been performed in this direction in recent years [1-5].

II. PROBLEM STATEMENT

The main goal of the original project is to digitalize all printed material accumulated over a long period of time in various research institutes in order to provide access to it for a larger number of readers and the possibility of subsequent automatic translation into the Kazakh language, including the use of the updated Latin alphabet. KazETU and all the research institutes involved have been linked by scientific and industrial ties for many years. Employees of the research institutes conduct classes at KazETU, and students of KazETU undergo scientific and industrial practice and even sometimes are invited to work at these research institutes.

In addition, the system is aimed at developing and

Waldemar Wójcik and Konrad Gromaszek are with Lublin University of Technology, Lublin, Poland. (e-mail: waldemar.wojcik@pollub.pl, k.gromaszek@pollub.pl).

Almas N. Temirbekov is with Al-Farabi Kazakh National University, Almaty, Kazakhstan (e-mail: almas.temirbekov@kaznu.kz).

Bakytzhan B. Omirzhanova is with Kazakh Research Institute of Processing and Food Industry, Almaty, Kazakhstan (e-mail: omirzhanova61@mail.ru).



implementing a software module for supporting research work, which will allow scientists and employees of research institutes to track their scientometric indicators in the main global bibliographic and abstract databases in real time. The relevance of creating this module is due to the fact that a fairly large part of the reporting of research institutes contains information on the publication activity and citation of scientists, rating of journals in which they are published. This information is especially important when preparing tender documents for grant funding, preparing documents for filling positions (including determining the members of the National Scientific Councils in priority areas of science and experts), and preparing quarterly and annual research reports institutions and accreditation. The processing of such data, as a rule, seems necessary in a short time, and taking into account the routine work of manually searching for data requires a lot of time. The developed module is also an analytical tool that allows evaluating the effectiveness and efficiency of the activities of research institutes and their scientists.

In fact, KazETU and all research institutes are a large Scientific and Educational Cluster. Therefore, despite the different departmental affiliation and different forms of ownership, complex automation of enterprises of the entire cluster within the framework of a single project looks quite logical.

The mention of the territorial proximity of the project participants in this case is not accidental, since the organization of high-quality communication channels within Kazakhstan, as in many developing countries, is still a significant problem. The lack of Internet access, fuzzy SLA, various types of blocking and high cost of traffic, unfortunately, is quite common. High-quality Internet access is not available throughout the country. This deprives cloud applications of a key component and the benefits of remote access. Therefore, direct physical L2 channels were used in the project to remotely connect participant campuses to centralized server resources. At the initial stage of the project, radio channels based on WiMax technology using UBNT radio frequency equipment at a frequency of 5.47 GHz were used to test the interaction. In the future, with a clear understanding of the project's economy, to increase the bandwidth of the channels used, it is planned to combine all the institutions involved in the project through fiber-optic communication channels.

In fact, the unique territorial proximity of the project participants, despite their departmental disunity and different ownership forms, made it possible to ensure the first steps towards the integration and centralization of their resources and further migration of part of the infrastructure to the cloud based on a commercial data center. At the same time, the use of closed physical channels in this project is dictated and justified, first of all, by the aspects of information security. This is due to the fact that the main project was working on building a structured database, some of which could be state secrets. Therefore, the issue of closed access was prioritized over such factors as usability, scalability, and others.

After the integration of disparate IT complexes of the project participants made it possible to centralize their IT infrastructure for a database of digitized printed resources on the basis of a

common server, it became possible to demonstrate the advantages of combining resources and start implementing cloud servers in other segments of the IT infrastructure of the entire scientific and educational cluster.

The main obvious advantages can be formulated as follows:

- the project participants, being structures that are not specialized in IT, get rid of non-core costs for the maintenance of computer equipment and its operation, while at the same time are getting modern technical tools and solutions at their disposal;
- thanks to infrastructure outsourcing, project participants are able to focus on research without looking at the capabilities of the IT infrastructure;
- project participants do not need to plan new costs for infrastructure modernization and development on their own.

These advantages of cloud infrastructure outsourcing have long been recognized by the global IT industry. In this regard, the work [6] is quite indicative, which provides an example of providing high-volume computing power on request for solving computational problems. Any user who connects to the capabilities described in the work gets the power of the supercomputer at their disposal.

A similar approach was used in the Computer Center for Collective Use (CCCU) of the Academy of Sciences of the Kazakh SSR in 1980-1985. Then, on the basis of the EU 1045 computer (analogous to the IBM System 370), the numerical solution of various computational problems was performed by the nearby research institutes (Institute of Mathematics, Metallurgy, Chemistry, and others). This experience is described in detail in [7].

Unfortunately, later, the use of personal computing resources based on X86 processors was widespread almost everywhere in Kazakhstan. After the liquidation of the CCCU, its experience in organizing remote access of third-party users to powerful computing resources was lost.

In [6], the main focus is on the use of virtualization technologies and the further use of IaaS, SaaS, and PaaS services in scientific work. It should be noted that for an ordinary user, a remote IT resource does not have any special differences, whether it is a physical or virtual computer. That is, if one gives a user terminal access to a multi-core supercomputer or to the program shell of this supercomputer, an unqualified user will not be able to feel the difference.

However, when creating a collective infrastructure with a limited budget, the factor of fault tolerance is important. Virtualization solutions are largely focused on fault tolerance, including the ability to meet disaster recovery requirements.

The city of Almaty, where the project is being implemented, is located in a zone of high seismic activity, so the risk of losing critical digital data is very high. In this regard, the disaster recovery requirement for the infrastructure is not superfluous. But it is necessary to understand that compliance with this requirement is associated with the creation of additional backup capacity outside the city. Such investments, taking into account the long distances in the Republic of Kazakhstan, can only become possible if the project is expanded and the infrastructure is distributed to a national scale using powerful computers.

As part of the current project to combine the infrastructure of KazETU and several research institutes to create a single distributed database of educational and scientific content obtained by digitizing the printed material accumulated over many years, a proposal arose among the project participants to test the methodology for using VDI (Virtual Desktop Interface) technology for the entire Scientific and Educational Cluster.

Testing and adaptation of technologies was carried out for one KazETU study group and several research institute workplaces involved in digitizing printed literature.

There are a lot of online resources on the Internet about organizing a server cluster and deploying a VDI shell. The site *itsave.ru* is notable for the fact that the company, being an authorized supplier of a number of manufacturers, provides a commercial estimate of the cost of deploying VDI infrastructure on hypervisors of several leading providers of such solutions: VMWare[®], CITRIX[®], Microsoft[®]. An approximate calculation of the server cluster for a complex of 100 and 500 VDI is also given. Of course, there are cheaper solutions on the market, such as Red Hat, Parallels, and even free KVM (Proxmox, described in [6]) and others. The project did not set out to compare these solutions, but only worked out the possibility of transferring a large group of users to VDI technology and demonstrating the capabilities of the approach as a whole.

Initially, the trial version of the ESXI server from VMWARE was used as the main shell (for 60 days). Generally speaking, the term VDI was first introduced to the market by this company, which owns more than 50% of the world market. It should be noted that this is the most expensive solution on the market. Therefore, in the future, one needs to either transfer the VDI shell of the scientific and educational cluster to another solution, or look for opportunities to reduce the cost of licenses in negotiations with the supplier.

As one of the options for testing works, we considered the possibility of using similar services from Amazon Web Services. This company is the undisputed global technology leader. However, given the fact that the company's servers are located outside the Republic of Kazakhstan, and some of the printed works of the research Institute that are translated into digital format may represent state secrets, the option of using a foreign cloud was initially rejected even at the stage of testing.

In order to determine the final technical specification of the terminal device, personal smartphones, tablets, and laptops of various modifications were used for the test group. Nuances of operation of the fulfilled specification can be recommended for mass use later. To visualize the virtual desktop, we used monitors of various configurations and MHL devices that provide communication between the smartphone and the monitor.

A cluster based on Supermicro[®] X86 servers was used as the hardware platform for the cloud being created. The data center was used as the data center *ahost.kz* one of the project participants stationed in Almaty. All WLL communication channels were consolidated from the project participants' buildings to the specified data center.

An additional advantage of this data center was that in the immediate vicinity there is an administrative office and an entry point to the Schoolnet, a network that unites 207 public schools

and 30 public colleges in the city with an audience of 220,000 students. By connecting Schoolnet to this cluster, one can get a huge breakthrough in the education of urban schoolchildren. But to provide the VDI infrastructure for such a large number of users, we need a fairly serious infusion of state scale.

The question of connecting other research institutes and universities to the created cluster remains open. For example, the inclusion of a supercomputer in the network, announced by the al-Farabi Kazakh National University, would strengthen the computing power of the cluster. There are no technological restrictions on creating such a cluster, but at this time departmental disunity and other local factors are hindering it.

An additional aspect considered in the framework of the described project of automation of the Scientific and Educational Cluster is the use of GaaS (Graphic as a Service) in scientific and educational work. When using the standard VDI shell, users are provided with video card resources with 512 KB video memory. This is extremely small for the current level of graphics technologies. The usual discrete graphics cards of middle range laptops have long passed the 1GB-2GB mark.

When using a video card with a small amount of video memory, the ability of VDI users to process various software applications is significantly limited, while for using graphics packages of application programs and game applications, video cards with significantly more memory are needed. To work with 3D graphics and animation, VDI needs the ability to connect a powerful enough video card.

As part of this project, a Supermicro video server was connected to an ESXI server with a VDI test infrastructure deployed for participants, equipped with a video card with the NVIDIA Kepler microarchitecture, which allows its video memory to be divided into several (usually four) VDI's.

At the time of writing, the used Grid K1 graphics card with 768 CUDA cores is significantly outdated. For example, in 2015, the Grid K1 video card was included in Azure Enterprise. Grid K2 video cards with 3000 CUDA cores produced by the same company NVIDIA and others appeared on the market.

The possibility of using the GaaS service in the educational process was tested in the framework of the described project. In the future, it will only be enough to scale this kind of power by installing several video cards with video memory expansion in one server, and later increasing the number of such video servers. This scheme is followed by the creators of some types of supercomputers, increasing the number of GPU cores. This same video memory, with a large number of GPU cores, can be used for rendering tasks and other graphics applications if the appropriate software is available.

The last point is very important, because the infrastructure being created (both using BYOD and VDI) is focused on the consumption of digital content. While using GaaS services, a Scientific and Educational Cluster can not only consume, but also create (generate) its own content. This task is constantly set by the state to the "digital" departments of the Republic of Kazakhstan. It is also mentioned in the latest state program "Digital Kazakhstan".

III. SOFTWARE PART OF A CLOUD PLATFORM

The selection of software underlying cloud storage of digitized work has required extensive research. This is due to the fact that research data from research institutes are not limited

to textual information, but also include data from satellites and sensors, geographical maps, field data, audio and video materials. Thus, the source of data stored in the cloud is not only scanned scientific works of archives and library collections, but also the results of equipment for remote sensing of the earth, precision farming, technology for visualizing space data on a map, quick access to highly detailed images for accurate mapping of field boundaries and crop rotation, fertility zones, as well as monitoring the state of vegetation. To work with such fragmented information, the system must have a flexible organization of storage of resources, provide the ability to effectively work with such information: splitting into collections, flexible search, and viewing (Fig. 1).

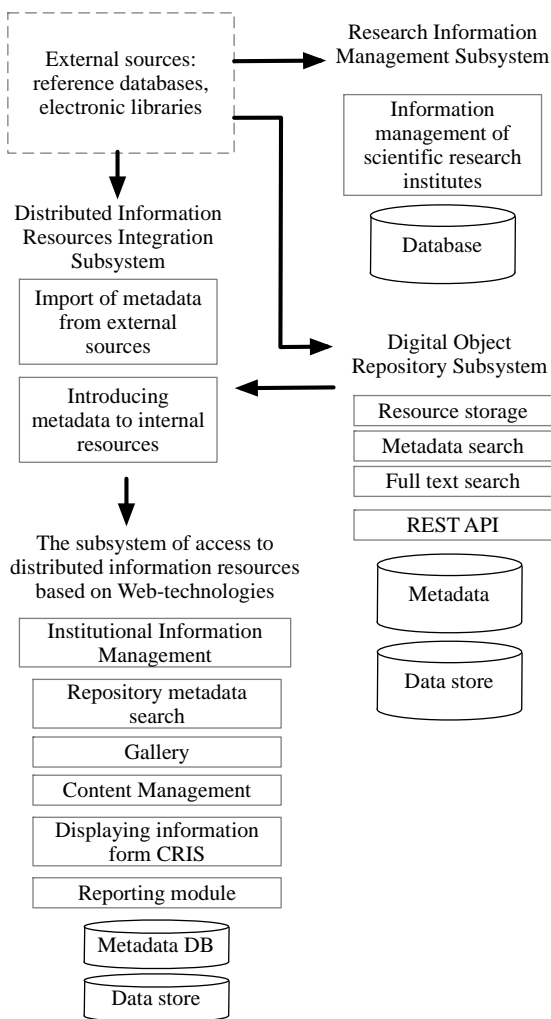


Fig. 1. Subsystems of the software underlying cloud storage.

Moreover, one of the priority requirements for this system was the ability to integrate distributed information resources based on standard protocols such as Z39.50, SRU, SRW. In addition, for integration with internal resources, the system must have an API. This requirement is associated with the need to import data from the own databases of some research institutes, as well as providing users with a single interface for finding resources and reporting.

The next requirement to the system is a full-text search of digitized text materials. Some repositories automatically extract

text when loading a new resource, provided that the text of this document has been previously recognized. This condition is quite difficult to comply with for most Eastern languages, since text recognition in these languages is implemented poorly enough in all software for optical character recognition.

In addition, one of the important requirements is a flexible system of user rights: the creation of user groups, the ability to specify the access of users of a given group to a given set of objects by the desired access method - loading, viewing, editing, deleting, changing attributes. In addition, the system must support centralized user authentication.

When choosing a subsystem of a repository of digital objects, ten institutional repositories and systems for creating electronic libraries were examined, and the experience of their use by various educational organizations and libraries was studied [8-14]. As a result of the analysis, the repository of digital objects DSpace was selected. DSpace supports more than 70 formats of information resources, and also has a more advanced system of user rights compared to the considered systems. When adapting the repository of digital objects, a comprehensive analysis of the available technical solutions was carried out, and the experience of adapting such systems was taken into account. The standard DSpace metadata scheme is adapted to the conditions accepted in the Republic of Kazakhstan and expanded with several fields

PostgreSQL database management system is used to store repository data. When developing the architecture of the information system, the possibility of using its cluster version of Postgres-XL, which allows you to store large amounts of data, increasing the reliability and availability of information through the use of replication mechanisms, as well as increasing the processing speed, was considered. Using sharding technology. From the user's point of view, Postgres-XL looks like a single database instance, that is, all client-side requests go through a standard connection. Architecturally, Postgres-XL consists of three types of components: global transaction monitor (GTM), coordinator, and data node.

However, in later versions of DSpace digital storage, metadata and content are stored in archive information packages, AIP, and the database is used as a data cache. According to the developers, this makes it easier to recover data in case of emergency, create backups, replicate data, make a checksum and / or digitally sign data. After analyzing the amount of information that DSpace stores in the database, it was concluded that the use of a cluster DBMS is not practical in the current environment.

The distributed information system ZooSPACE developed at the Institute of Computational Technologies SB RAS [15,16] was chosen as integrating software. It combines data from various sources of information, providing access to heterogeneous distributed information in accordance with standard protocols (SRW/SRU, Z39.50). The system works on the basis of original ZooPARK-ZS and LDAP servers and Apache web server, providing end-to-end information search in heterogeneous databases, information extraction in standard schemes and formats and its display.

A web application has been developed to provide a standardized single user interface for all functions and modules that make up the distributed information system. The main objectives of this subsystem are:

- providing detailed information about the scientific directions, activities, the most significant scientific achievements, and services of the research institutes;
- information about their employees including main scientific contribution, references to basic scientometric databases;
- bibliometric indicators of scientific productivity of employees of research institutes;
- the list of their digitized publications;
- flexible search both in the repository of digital objects, and the main scientometric databases.

The web application is developed using the Django web framework and runs on the Gunicorn WSGI HTTP server with the nginx HTTP server installed as a reverse proxy server. The scheme of interaction of the web portal with the components of a distributed information system is shown in Fig. 2.

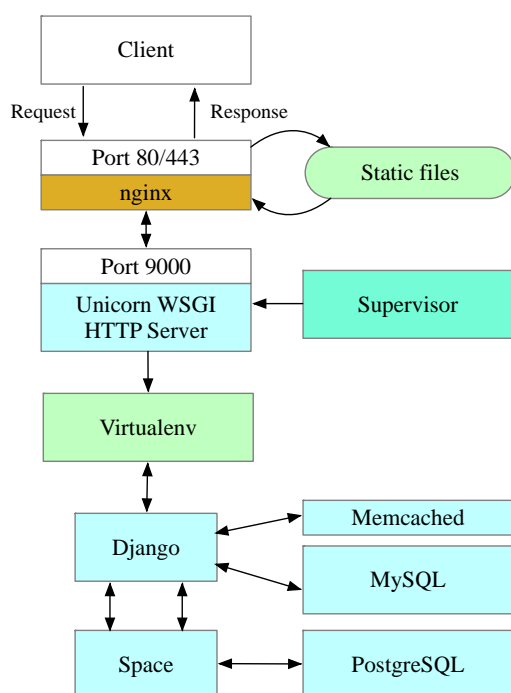


Fig. 2. Scheme of interaction of the web application with components of the cloud software.

The architecture of the web application allows to extend its functionality through modules. In order to provide the user with a single user interface to all the functions of the integrated distributed information system, the integration of the DSpace digital object repository and the web application was carried out. In particular, it is possible to search for data in the repository from the web application. It is implemented based on the use of the REST DSpace API, which provides a programming interface for communities, collections, item metadata and files. An HTTP request is made to DSpace REST using the curl utility on each search query. An unprivileged user has been created in DSpace on behalf of which the search is performed. Responses from DSpace REST are returned in JSON format, which are parsed in the developed Integrator module. Substantial attention is paid to providing flexibility and convenience in finding information resources. The search is successfully performed taking into account the various endings of the searched word. It has been revealed from the guide to

DSpace that the possibility of specifying different spellings of the scientist name is not implemented at the moment. In the developed module, various spelling of his name is taken into account, including in different languages when searching for resources by the author. An example of the research results is shown in Fig. 3.

Numerical solution of the problem of optimizing the process of oil displacement by steam
 Temirbekov, N. M.; Baigereyev, D. R.
 AIP Publishing (2016)

Study of the Difference Scheme for Model Problem of the Gaslit Process
 Temirbekov, Nurlan; Turarov, Amankeldy
 International Conference "Functional Analysis in Interdisciplinary Applications (2017)

Построение научно-образовательного кластера на облачных вычислительных технологиях
 Темирбеков, Н.М.; Байгереев, Д.Р.; Темирбеков, А.Н.; Исакова, А.М.; Омиржанова, Б.Б.
 Берел, 2018

Fig. 3. An example of search results for information resources by author, taking into account the different spelling of his name.

The software implementation of the module of integration and support of research work is based on the interaction of the web application with the scientometric database API with further analysis and systematization of data. For this, a table was created containing the digital identifiers of each employee of the research institute:

- Scopus Author ID;
- Web of Science Researcher ID;
- Google Scholar Citation Profile;
- ORCID.

These identifiers are currently filled in by web application content managers. The main functionalities of the created module include:

- obtaining a list of publications of employees of research institutes;
- synchronization of information with profiles of scientists on the website of the Academgorodok;
- display of publications of a research institute;
- displaying the number of links to publications and calculating the h-index of the scientist, taking into account and excluding self-citation;
- reporting, reflecting the assessment of the effectiveness and efficiency of research institutes, the conclusion of structured current and comparative data on scientometric indicators of employees. This feature takes into account reporting periods for benchmarking by quarter, year, and five-year breakdown;
- exporting data to various formats.

It should be noted that the h-index is not returned on request. Its calculation is implemented based on an analysis of the list of publications and their citations.

In general, the developed module saved employees of research institutes from the routine of manual work on extracting and analyzing information from scientometric databases.

CONCLUSION

The article analyzes the process of developing a scientific and educational cluster to support the research work of a number of institutions located in Almaty. The results of the analysis of existing technical solutions for creating systems of this class are presented and the technical details of their implementation are described. The architecture of the software part of a distributed information system is presented.

In general, it can be noted that the developed system fully provides the necessary computing resources for ongoing research and educational processes, simplifying the prospect of its further development, and allows you to build an advanced IT infrastructure for managing intellectual capital, an electronic library in which all books and scientific works of the Kazakhstan Engineering and Technological University and research institutes of the Akademgorodok in Almaty.

According to the results of the project, it can be clearly stated that the introduction of a system to support the scientific and educational cluster in the educational and research work of universities and research institutes does not present great difficulties and has great prospects from the technical side in comparison with the use of a large number of personal computers.

REFERENCES

- [1] S. Khode, S. S. Chandel, "Adoption of open source software in India," *DESIDOC Journal of Library and Information Technology*, no. (35) 1, pp. 30-40, 2015.
- [2] M. Termens, M. Ribera, A. Locher, "An analysis of file format control in institutional repositories," *Library Hi Tech*, no. (33) 2, pp. 162-174, 2015.
- [3] K. Riddle, "Creating policies for library publishing in an institutional repository: Exploring purpose, scope, and the library's role," *OCLC Systems and Services*, no.31 (2), pp. 59-68, 2015.
- [4] A. Abrizah, M. Hilmi, N. A. Kassim, "Resource-sharing through an inter-institutional repository: Motivations and resistance of library and information science scholars," *Electronic Library*, no. 33 (4), pp. 730-748, 2015.
- [5] R. M. Marsh, "The role of institutional repositories in developing the communication of scholarly research," *OCLC Systems and Services*, no. 31 (4), pp. 163-195, 2015.
- [6] A. Baranov and E. Kiselev, "Cloud services for scientific high-performance computing based on the Proxmox platform," *Computational technologies*, vol. 24, no. 6, 2019, pp.5-12.
- [7] I. T. Pak, "From the history of Informatics development in Kazakhstan," Almaty, 2012.
- [8] H. Franchke, J. Gamalielsson, and B. Lundell, "Institutional repositories as infrastructures for long-term preservations," *Information Research*, vol. 22 (2), no. 757, pp. 1-27, 2016.
- [9] C. Hippenhammer, "Comparing institutional repository software: pampering metadata uploaders," *The Christian Librarian*, vol. 59, no. 1, pp. 1-6, 2016.
- [10] K. Baughman, "Institutional repositories in the Czech republic," *Gleeson Library Librarians Research*, vol. 10, pp. 1-29, 2016.
- [11] M. N. Ravikumar and T. Ramanan, "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*, vol. 18, no. 2, pp. 76-90, 2014.
- [12] M. Castagné, "Institutional repository software comparison: DSpace, ePrints, Digital Commons, Islandora and Hydra (Report)", University of British Columbia, 2013.
- [13] R. Cullen and B. Chawner, "Institutional repositories in New Zealand: comparing institutional strategies for digital preservation and discovery," *Proceedings of the IATUL Conference*, vol. 18, pp. 1-11, 2008.
- [14] M. Grzesińska, M. Waszczyńska, and B. Pańczyk, "JEE database applications performance," *Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska*, vol. 6, no. 4, pp. 73-76, 2016.
- [15] O. L. Zhizhimov, A. M. Fedotov, and O. A. Fedotova, "Building a typical model of an information system for working with documents on scientific heritage", *Vestnik NGU. Informacionnye tehnologii*, vol. 10, no 3, pp. 5-14, 2012.
- [16] Yu. I. Shokin, A. M. Fedotov, O. L. Zhizhimov, and O. A. Fedotova, "The control system of electronic libraries in IRIS SB RAS", *4th All-Russian symposium, Saint Petersburg*, October 6-8, 2014.