Proceedings of International Mathematical Sciences
ISSN:2717-6355, URL: https://dergipark.org.tr/tr/pub/pims
Volume III Issue 1 (2021), Pages 32-37.
DOI: https://doi.org/10.47086/pims.613534

DEVELOPMENT OF AN INFORMATION SYSTEM FOR STORING DIGITIZED WORKS OF THE ALMATY ACADEMGORODOK RESEARCH INSTITUTES

NURLAN TEMIRBEKOV*, DOSSAN BAIGEREYEV**, ALMAS TEMIRBEKOV***, AND BAKYTZHAN OMIRZHANOVA****
*KAZAKHSTAN ENGINEERING TECHNOLOGICAL UNIVERSITY, ALMATY, KAZAKHSTAN,
**S. AMANZHOLOV EAST KAZAKHSTAN STATE UNIVERSITY, UST-KAMENOGORSK, KAZAKHSTAN, +77053312023
***AL-FARABI NATIONAL UNIVERSITY, ALMATY, KAZAKHSTAN,
****KAZAKH RESEARCH INSTITUTE OF PROCESSING AND FOOD INDUSTRY, ALMATY, KAZAKHSTAN

Abstract. The present article describes the architecture of the integrated distributed information system created for storing digitized works of employees of Almaty Akademgorodok research institutes (Kazakhstan) and providing access to them using Web technology. Comparative analysis of two data storage systems for storing digitized works, Ceph and GlusterFS, is provided. The description of the software part of the information system is provided which consists of four subsystems: repository of digital objects, subsystem for managing current research information, subsystem of integration of distributed information resources, subsystem of access to distributed information resources based on Web technologies. The relation between the subsystems and their integration is described. The work defines the requirements to the repository of digital objects. The requirements for the repository of digital objects are defined; a comparative analysis of open source software used for these purposes is made.

1. Introduction

For decades, scientists of Almaty Academgorodok research institutes have been conducting enormous research in leading areas of the agro-industrial, processing, microbiological, seismological and other areas producing hundreds of thousands of articles, technical reports, and other documents. The latter also include digital research materials in the form of statistical, cartographic, multimedia data which are derived using radars, telescopes, and satellites. However, it should be recognized that the results of these studies remain inaccessible to the vast majority of

2020 *Mathematics Subject Classification.* 68M14, 68-04, 68-06.
Key words and phrases. integrated distributed information system; data storage; Ceph, GlusterFS, digital object repository; DSpace; CRIS; web portal.

Submitted on 31.08.2019, Accepted on 10.07.2021.
Communicated by Ahmet Mesut Razbonyali.

researchers and employees of the agro-industrial complex in the age of the information explosion. One of the reasons for this problem is the lack of a publicly available single repository of information and knowledge base in the field of agriculture. In addition, important works created more than half a century ago and stored in the archives of libraries in paper form take an unpresentable form over the years.

For this reason, a team of scientists of Kazakhstan Engineering Technological University and Academset LLP has created an integrated distributed information system of Academgorodok, acagor.kz, the main objectives of which are: (1) providing reliable storage of the results of intellectual creative activity of employees of research institutes including digitized works, geographic materials (maps, satellite images, field observations), audio and video recordings; (2) management of current research information; (3) external and internal integration of information resources; (4) providing a single user interface for all functions and modules of a distributed information system, providing a "transparent“ search and user access to documents; (5) optical character recognition of handwritten manuscripts created in past centuries based on neural networks.

Therefore, the objective of the work is not only to preserve the rich heritage of the research institutes, but also to provide access to them and the ability to quickly search for the necessary information. The creation of such a specialized information resource could not only act as a reservoir of valuable research results, but also unite employees of research institutes, agricultural sector workers and other users working in this field.

This article presents a description of the information system for the implementation of the above tasks. The structure of the article is as follows. Section II provides the description of the data storage system. Section III provides the description of the software part of the information system. Section IV outlines plans for further work on the modernization of the information system.

2. Data Storage

At the initial stage, a NAS type architecture was chosen when conducting tests on a relatively small amount of data. However, the exponential growth of stored information led to a revision of the data warehouse architecture.

To organize the storage of data, it was decided to use distributed file systems. The choice of a distributed file system is made on the basis of compliance with the following criteria: (1) high reliability of storage; (2) high availability of data; (3) fault tolerance; (4) decentralization; (5) scalability; (6) low unit cost of storage; (7) ease of deployment and operation.

Ceph [1] and GlusterFS [2] were considered as data storage systems. Both of the systems provide high performance and scalability. However, these systems are architecturally opposite.

GlusterFS works in user space using FUSE technology, therefore it does not require support from the operating system kernel and works on top of existing file systems. The strengths of GlusterFS include simpler deployment. Unlike Ceph, GlusterFS does not require a separate server to store metadata, it is stored along with the data in the extended file attributes. Due to the lack of binding to the centralized meta-data server, the file system provides almost unlimited scalability. However, GlusterFS has fewer options and is less flexible compared to Ceph.

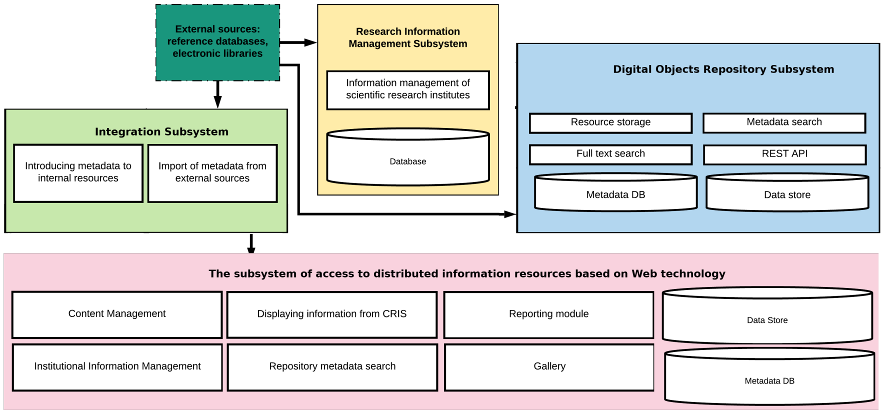
External sources: reference databases, electronic libraries
Research Information Management Subsystem
Information management of scientific research institutes
Database
Digital Objects Repository Subsystem
Resource storage
Metadata search
Full text search
REST API
Metadata DB
Data store
Integration Subsystem
Introducing metadata to internal resources
Import of metadata from external sources
The subsystem of access to distributed information resources based on Web technology
Content Management
Displaying information from CRIS
Reporting module
Data Store
Institutional Information Management
Repository metadata search
Gallery
Metadata DB

Figure 1. Relation between subsystems of the information system.

Unfortunately, our tests conducted on the basis of three servers showed a relatively slow work of the FUSE driver when working with small files, as well as poor responsiveness when rebuilding. Testing of the file system was carried out on the basis of servers with the following characteristics: Intel Xeon E5 2V Core, 8 GB DDR3 ECC, 500 GB HDD RAID 5.

The advantage of Ceph over GlusterFS is the lack of single points of failure and almost zero maintenance costs for recovery operations. In Ceph, the data array is automatically rebalanced when adding or removing new nodes. It occurs almost imperceptibly to clients and ensures a high survivability of the system.

According to the developers, the current version of CephFS is not stable for production use, but the test, deployed on the basis of three physical nodes and one virtual machine, showed no problems. Four servers were used in the test data warehouse: an administrative node (hub.acagor.kz) and three servers with data (node1.acagor.kz, node2.acagor.kz, node3.acagor.kz). Testing was performed within one month with the emergency shutdown of one of the servers. In this case, the rebalancing of the cluster occurs without a second downtime and is transparent to clients.

As a result of numerous studies, preference is given to Ceph object storage network. Ceph showed more acceptable results in terms of performance compared to GlusterFS when working with a digital repository of DSpace (see below) due to its assets storage features.

3. Architecture of the Information System

The information system comprises the following subsystems which are presented in Fig. 1: (1) Subsystem of repository of digital objects, (2) Subsystem for managing current research information; (3) Subsystem of integration of distributed information resources; (4) Subsystem of access to distributed information resources based on web technologies. Detailed information about the subsystems is given below.

3.1. **Subsystem of Repository of Digital Objects.** The first subsystem is intended for long-term storage of information described in the Introduction. The following system requirements were defined to the software underlying the subsystem: (1) the ability to store various types of resources including images, maps, audio and video recordings; (2) flexible storage organization: possibility of arbitrary grouping of resources by various criteria; (3) the ability to authenticate users and manage user roles through LDAP; (4) the ability to access external scientometric databases (i.e. Web of Science, Scopus etc.); (5) the ability to integrate with internal resources through APIs; (6) text recognition and full-text search; (7) open source software.

The capabilities of many software and technical solutions were analyzed including Ambra, Digital Commons, DSpace, ePrints, Evergreen ILS, Greenstone, Fedora Commons, Invenio, RODA, and VuFind. The experience of using each of these systems by scientists from various countries was also studied [3, 4, 5, 6, 7, 8]. Analyzing the advantages and disadvantages of the listed systems, Greenstone, ePrints, and DSpace 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 uploaded. However, this system supports only a limited number of formats. Storage of geographical maps, as well as other results of scientific activities that have a more complex structure is not provided. ePrints supports more metadata formats, but does not support the extended Dublin kernel. The system supports various user roles. The strengths of DSpace include a more sophisticated system of user rights compared to the systems considered: various research institutes can have their own areas within the system. In each institute, certain employees that are responsible for pre-moderation may be appointed. 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.

As a result of the analysis, DSpace was chosen as a underlying subsystem. However it did not fully satisfy some of the requirements. Changes were made to its configuration during its compilation 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 the Committee for the Control of Education and Science of the Republic of Kazakhstan", "Full bibliographic reference in accordance with state standard" and others. In addition, reporting subsystem did not cover the requirements.

To store the repository data, the PostgreSQL database management system is used. In developing the information system architecture, the possibility of using its cluster version, Posgtres-XL, was considered. However, in later versions of the DSpace digital repository, metadata and content are stored in archival information packets, AIPs, and the database is used as a data cache. After analyzing the amount of information that DSpace stores in the database, it was concluded that the use of cluster DBMS is impractical under current conditions.

3.2. **Current Research Information Management (CRIS) Subsystem.** A literature review showed that there are few CRIS systems that can be integrated to DSpace. There is a powerful system created for this purpose at the Institute of Computational Technologies of the Siberian Branch of Russian Academy of Science (SB RAS) [9].

As a result of the study, an extension of the DSpace system, DSpace-CRIS was chosen as a research management system. This system allows to store information on research organizations, information about the employee of research organizations, various spellings of researcher's 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 is integrated with a DSpace instance, which allows to view the publication of scientists. The CRIS-system allows to export information about the publications of the scientist in popular formats.

3.3. **The Subsystem of Access to Distributed Information Resources Based on Web Technologies.** This subsystem is designed to provide a single user interface for all functions and modules included in the distributed information system. To date, it successfully solves the following tasks: (1) providing detailed information about the activities of research institutes and their employees; (2) flexible search both in the repository of digital objects and the external databases; (3) the opportunity to discuss topical issues on the forum, providing users the opportunity to share their work with colleagues; (4) organization of conferences including submissions of participants' papers (and storing in the repository), sending the papers to reviewers, booking a hotel, etc. The subsystem was 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.

3.4. **Integration Subsystem.** As integrating software, the ZooSPACE distributed information system was chosen [9]. The literature review did not reveal more suitable software for this purpose. 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. To implement search in the digital object repository, a web portal was integrated with DSpace using the DSpace REST API, which provides a programming interface to communities, collections, item metadata and files. As a result of the integration metadata and links to materials uploaded to the repository subsystem are displayed on the scientist's profile page and on the information page of research institutes. Search by metadata is available. Additionally, filtering by keywords, institutes, date and language of publication was implemented.

4. Conclusion

The developed information 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.

Currently, the authors of the papers are working on full-text search in the repository of digital objects. Analysis of existing optical character recognition software (including proprietary) for digitization of texts revealed significant difficulties. This is partly due to the quality of the source materials, as well as the difficulty of recognizing characters of some Asian alphabets. In addition, work is underway on the use of neural networks for the recognition of manuscripts and old printed texts. The results of the research in this area will be presented in future papers.

Acknowledgments. This research was supported by Science Committee of Ministry of Education and Science of the Republic of Kazakhstan (grant No. AP05131806).

References

[1] A. D'Atri, and V. Bhembre, and K. Singh, Learning Ceph - Second Edition: Unifed, scalable, and reliable open source storage solution (2017).

[2] I. T. Avery, Glusterfs (2012).

[3] H. Franchke, and J. Gamalielsson, and B. Lundell, Institutional repositories as infrastructures for long-term preservations, Information Research 22, nr 757 (2016) 1-27.

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

[5] K. Baughman McDowell, Institutional repositories in the Czech republic, Gleeson Library Librarians Research 10 (2016) 1-29.

[6] 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 18, nr 2 (2014) 76–90.

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

[8] R. Cullen, and B. Chawner, Institutional repositories in New Zealand: comparing institutional strategies for digital preservation and discovery, Proceedings of the IATUL Conference 18 (2008) 1-11.

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

Nurlan Temirbekov,
93A, al-Farabi ave., 050060 Almaty, Kazakhstan, Phone: +77772794876
Email address: temirbekov@rambler.ru

Dossan Baigereyev,
34, 30th Guardian Division str., 070000 Ust-Kamenogorsk, Kazakhstan, Phone: +77053312023
Email address: dbaigereyev@gmail.com

Almas Temirbekov,
71, al-Farabi ave., 050040 Almaty, Kazakhstan, Phone: +77052500653
Email address: almas_tem@mail.ru

Bakytzhan Omirzhanova,
238G, Gagarin str., 050060 Almaty, Kazakhstan, Phone: +77770645051
Email address: omirzhanova61@mail.ru