Journal of Engineering Science and Technology Review
Special Issue on Telecommunications, Informatics, Energy and Management 2019

Conference Article

JOURNAL OF
Engineering Science and Technology Review

www.jestr.org

Development of a Distributed Information System of the Almaty Academgorodok

Nurlan Temirbekov[1], Dossan Baigereyev[2], Almas Temirbekov [3] and Bakytzhan Omirzhanova[4]

[1]Kazakhstan Engineering Technological Universiy, Almaty, Kazakhstan,
[2]East Kazakhstan State Technical University, Ust-Kamenogorsk, Kazakhstan,
[3]Al-Farabi Kazakh National University Almaty, Kazakhstan,
[4]Kazakh Research Institute of Processing and Food Industry, Almaty, Kazakhstan,

Received 25 September 2019; Accepted 12 February 2020

Abstract

The present article describes the architecture of an integrated distributed information system used to store and manage digitized works of employees of research institutes of the Almaty Academgorodok. The Ceph open source software object storage network is used as data storage. Testing of the POSIX-compatible CephFS file system abstraction is performed. The software part of the information system 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 description of the software part of the information system is provided. Integration between the subsystems of the information system is performed.

Keywords: integrated distributed information system; data storage; digital object repository

1. Introduction

For several decades, research and development institutes located in the Almaty Academgorodok (Kazakhstan) have carried out important and extensive scientific research in leading areas of the agro-industrial, processing, microbiological, seismological and other areas. The results of their intellectual work are published in the form of technical reports, monographs, and articles. However, it should be recognized that in the age of the information explosion, the results of these studies remain inaccessible to the overwhelming majority of researchers. 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.

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 this regard, the objective is not only to preserve the rich heritage of research institutes, but also to provide access to them and the ability to quickly search for the necessary information.

To this end, a team of scientists from Kazakhstan Engineering Technology University and Academset LLP has created an integrated distributed information system of Academgorodok, acagor.kz, the main tasks of which are:

- reliable storage and management of large data, including digitized works of scientists of scientific research institutes, geographic materials (maps, satellite images, field observations), audio and video recordings, etc.;
- implementation of a full cycle of digital content formation;
- management of current research information;
- external and internal integration of information resources;
- providing a single user interface for all functions and modules that are part of a distributed information system, providing a "transparent" search and user access to documents, both for review and for the analysis of the facts contained in them.

Currently, the information system successfully performs most of the tasks. This article presents a description of the information system for the implementation of the above tasks.

2. Data storage

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

To organize data storage, it was decided to use a distributed file system. The choice of a distributed file system is made on the basis of compliance with the following criteria:

- high reliability of storage;
- high availability of data;
- fault tolerance;
- decentralization;

*E-mail address: temirbekov@rambler.ru

ISSN: 1791-2377 © 2020 School of Science, IHU.
All rights reserved.

- scalability;
- low unit cost of storage;
- ease of deployment and operation.

As a result of numerous studies, preference is given to Ceph open source software object storage network [1]. This network is a highly scalable petabyte repository that can host thousands of nodes [2]. The built-in mechanisms of duplicated data replication ensure a high survivability of the system; when adding or removing new nodes, the data array is automatically rebalanced with changes.

Ceph provides a choice of three different abstractions for working with storage: object storage abstraction (RADOS Gateway), block device abstraction (RADOS Block Device) and POSIX-compatible file system abstraction (CephFS). 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). All servers and connected clients are managed from the administrative node, and there are monitors on the other servers, a block device for storing data and a hash of sums. 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.

3. The software part of the information system

The software part of the distributed information system consists of the following subsystems depicted in Fig. 1:

-Digital objects repository subsystem;

-Subsystem for managing current research information;

-Subsystem of integration of distributed information resources;

-Subsystem of access to distributed information resources based on web technology.

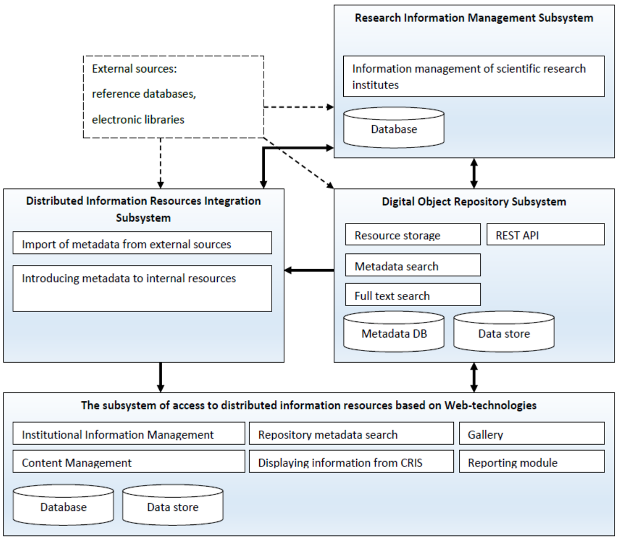
Research Information Management Subsystem
Information management of scientific research institutes
Database
External sources:
reference databases,
electronic libraries
Distributed Information Resources Integration Subsystem
Import of metadata from external sources
Introducing metadata to internal resources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

Fig 1. Architecture of the software part of the information system

3.1 Digital Object Repository

The Digital Object Repository subsystem is intended for long-term storage of the results of scientific research institutes. The following basic requirements were put forward to the software underlying the subsystem:

1. Ability to work with documents of arbitrary formats (for example, geographical maps, audio and video materials);
2. Flexible organization of resource storage;
3. Flexible user rights: creating user groups; the ability to specify the users access 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;
4. The ability to integrate distributed information resources based on standard Z39.50/SRU/SRW protocols;
5. The presence of a software interface for integration with internal resources;
6. Optical character recognition of 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: Ambra, Digital Commons, DSpace, ePrints, Evergreen ILS, Greenstone, Fedora Commons, Invenio, RODA, and VuFind. Furthermore, the experience of their use in organizations in New Zealand, the Czech Republic, Sri Lanka and other countries were studied [3-8].

Analyzing the advantages and disadvantages of these systems, DSpace repository of digital objects was chosen. When assembling DSpace, changes were made to its configuration in order to adapt to the conditions established in the Republic of Kazakhstan.

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, which allows storing large amounts of data, increasing the reliability and availability of information through the use of replication mechanisms, as well as increasing processing speed using sharding technology. However, in later versions of DSpace, metadata and content are stored in archival information packets, AIPs, and the database is used as a data cache. According to the developers, this allows to more easily restore data in case of emergency, to make backups, replicate data, make 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 cluster DBMS is impractical under current conditions.

The internal organization of storage of resources in the DSpace system is structured as follows: 7 thematic communities 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.).

3.2 Subsystem for managing current research information

An extension of DSpace, DSpace-CRIS was chosen as a research management system. The system allows to store information on research organizations, employees of research organizations, various spellings of the researcher's name, 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 technology

This subsystem is designed to provide a standardized uniform user interface for all functions and modules included in the distributed information system.

The main objectives of the subsystem are:

- providing detailed information about the activities of research institutes, information about their employees and the list of digitized works;
- flexible search both in the repository of digital objects, and the scientometric databases.

The subsystem is designed using the Django web framework. The architecture of the web portal allows to extend its functionality through modules. In addition to the listed features, the several additional modules are implemented. For example, the conference module, in which the full cycle of work on the organization of conferences is carried out. It provides accepting materials from conference participants with notification of the change in application status to booking a hotel. This module was tested in the organization of the International Conference "Science, Production, Business: Current State and Ways of Innovative Development of the Agrarian Sector by the Example of the Baiserke-Agro Agricultural Holding" held in April 2019.

3.4 Integration subsystem of the integrated information resources

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 [9-11].

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. DSpace REST API is deployed as a standard Web application; an unprivileged "restuser" user has been created in DSpace on whose behalf search queries are made. All requests are made using the curl utility; JSON responses are analyzed in the Integrator web portal module.

4. Conclusion

Currently, the constructed information system is used to store digitized articles, monographs and technical reports of the Institute of General Genetics and Cytology, the Institute of Human and Animal Physiology, the Kazakh Research Institute of Processing and Food Industry, the Kazakh Research Institute of Fruit and Viticulture, the Kazakh Scientific Research Institute of Soil Science and Agrochemistry named after U. Uspanov, Kazakhstan Engineering and Technological University, Research Institute of Microbiology and Virology.

As a result of the integration of the web portal with the repository of digital objects, metadata and links to materials uploaded to the storage subsystem are displayed on the profile page of the scientist and on the information page of research institutes. Search by metadata and full-text search is available. In addition, filtering by keywords, institutions, date and language of publication is implemented. Currently, more than 700 works are stored in the repository. The earliest publication dates back to 1898.

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.

References

1. https://ceph.com/.
2. S. A. Weil, S. A. Brandt, L. Miller, D. E. Darrell, and C. Maltzahn,"Ceph: A Scalable, High-performance Distributed File System" in OSDI '06. Proceedings of the 7th Symposium on Operating Systems Design and Implementation, pp. 307-320, Seattle, Washington: USENIX Association, 2006
3. H. Franchke, J. Gamalielsson, and B. Lundell, "Institutional repositories as infrastructures for long-term preservations," Information Research, vol. 22, nr 757, pp. 1-27, 2016
4. C. Hippenhammer, "Comparing institutional repository software: pampering metadata uploaders," The Christian Librarian, vol. 59, nr 1, pp. 1-6, 2016
5. K. Baughman McDowell, "Institutional repositories in the Czech republic," Gleeson Library Librarians Research, vol. 10, pp. 1-29, 2016
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, vol. 18, nr 2, pp. 76–90, 2014
7. M. Castagné, "Institutional repository software comparison: DSpace, ePrints," Digital Commons, Islandora and Hydra (Report), University of British Columbia, 15 p., 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, pp. 1-11, 2008
9. 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," Bulletin of the NSU. Information Technology, vol. 10, nr 3, pp. 5-14, 2012
10. Yu. I. Shokin, A. M. Fedotov, O. L. Zhizhimov, and O. A. Fedotova, "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, vol. 1, pp. 11-39, 2014

11. O. L. Zhizhimov, A. M. Fedotov, and Yu. I. Shokin, "Technological platform for mass integration of heterogeneous data," Bulletin of the Novosibirsk State University. Series: Information Technology, 11, nr 1, pp. 24-41, 2013