A New Version of the Software for the Information System "Scientific Heritage of Russia"

Konstantin Pogorelko[0000-0002-7598-8813]

Joint Supercomputer Center of Russian Academy of Sciences – Branch of Federal State Institution "Scientific Research Institute for System Analysis" of Russian Academy of Sciences, Leninskiy pr., 32a, 119334, Moscow, Russia
konstpog@yandex.ru

Abstract. The information system "Scientific Heritage of Russia" has been created in stages since 2007. Currently, the existing software does not meet the needs of the system and complicates its further development. It was decided to implement the new version of the software in the asp.net core cross-platform environment. The article describes the decisions made in the implementation of software and modernization of the data structure. Particular attention is paid to the development of information retrieval tools.

Keywords[1]**:** databases, information systems, databases, e-library, digital library, software, information retrieval, asp.net core.

The Electronic Library (DL) "Scientific Heritage of Russia" began to be created in 2007 [1] and functions in full for over 10 years [2]. The library is in constant demand [3]. Distribution by year quantity requests for electronic publications included in the DL, is shown in Fig. 1, and number of requested pages is shown on Fig. 2.

The software supporting the library has been modified several times. So, in 2013, the presentation of full text publications was separated into a separate subsystem [4].

The current state of the library is given in [5].

Nowadays there is a need to rework the software. This is due to both the increased requirements for the search capabilities of the system and its speed and for a more complete satisfaction of the requirements put forward by the technological processes of information preparation.

Since the development tools on which the existing system is based are outdated, it was decided to radically redesign both the software that supports the system and its internal structure and interfaces.

---

[1] CDSSK–2020: International Conference "Common Digital Space of Scientific Knowledge", November 10–12, 2020, Moscow, Russia

EMAIL: konstpog@yandex.ru (Konstantin Pogorelko)
ORCID: 0000-0002-7598-8813 (Konstantin Pogorelko)

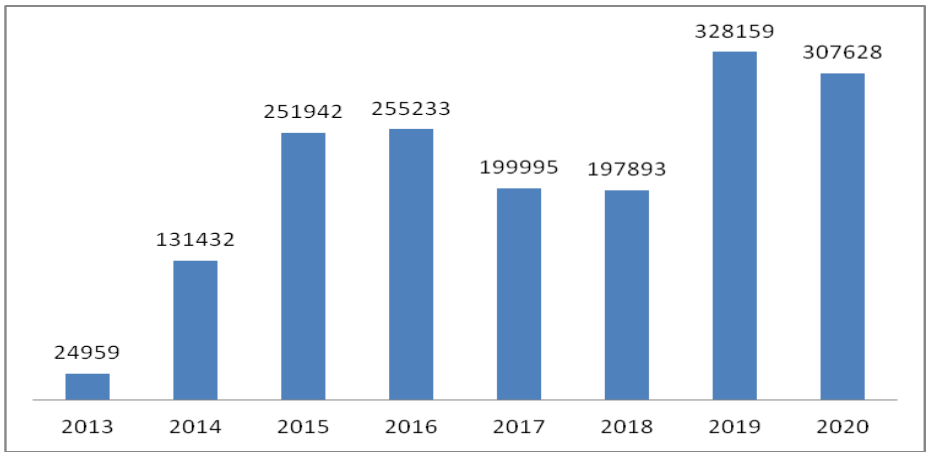

Fig. 1. Number of requests for electronic publications by year.

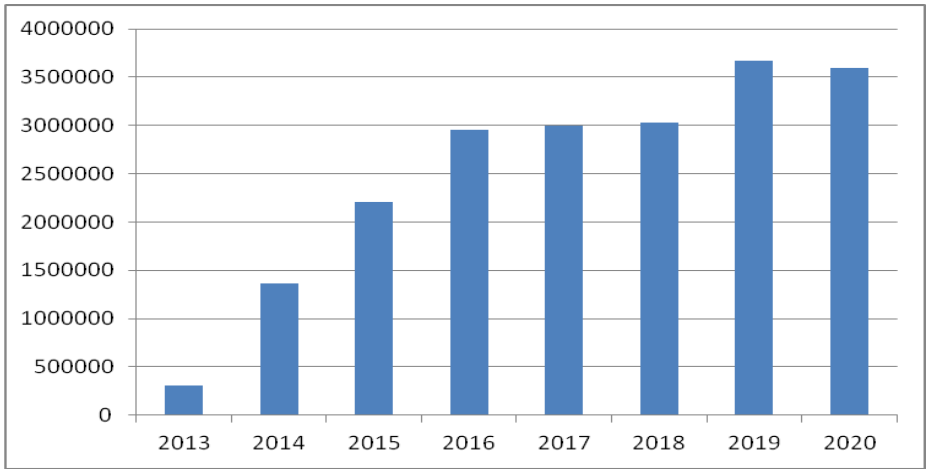

Fig. 2. Number of requested pages by year.

To implement the new version of the software, a modern cross-platform asp.net core environment was chosen. The main feature of this platform is the versatility of applications developed on its basis. These applications can run on both Microsoft and Unix operating systems. In addition, when deploying web applications developed on the basis of this platform, they are formatted as an executable file, which eliminates the process of interpretation and ensures their high efficiency. For storing data, the SQL server PostgreSQL was chosen, which has implementations for different platforms and provides fairly effective data manipulation.

The system being developed is combined into a single project, consisting of the following parts: a subsystem for preparing meta information for various objects, a subsystem for preparing electronic publications and a subsystem for information search and

visualization. This solution allows you to take into account possible changes in the composition and data structure at once in the entire project. In addition, depending on the performance of the servers and the load on the system, it becomes possible to deploy the system both on one server and to distribute the components of the project to different servers.

In previous versions of the system, information was located in several databases, between which the exchange protocols were organized. In addition to publishing delays and increasing the likelihood of failure, this created certain difficulties in organizing cross-references between objects that were placed in different databases. In the new version of the system, all data is contained in a single database, which ensures uniformity of their processing, facilitates data manipulation and ensures that there are no delays in their publication.

To ensure quality control of the input data, a two-level system of authority has been adopted. At the first level, the operator, who has the authority to enter data, enters the data. Upon completion of the input, the operator sets the sign "input completed". The issuing editor, reviewing documents with the sign of completion of the entry, checks the entered data and, if the data was entered with the required quality, sets the sign "published". After the document receives this status, it becomes available for search and display in the library. The operator who entered the document that received the "published" status loses the right to correct this document if he does not have the rights of the issuing editor.

The system implements tools to support procedures for identifying and processing erroneous data. Errors can be detected both in manual mode when viewing data, and in automatic mode, for example, when checking the correctness of external links. All detected errors are assigned in the form of a list to the object to which they refer. The operator has the ability to view the identified errors and, after correcting them, assign the appropriate status to the document.

In the structure of the database, it is possible to supply the links between the metadata of various objects with additional attributes that concretize this link. For example, for the relationship of objects "person" – "publication", you can specify the value of the attribute "author", "editor", "about him", etc.

The new version of the system has significantly expanded its search capabilities. The search module of the system allows you to process queries consisting of an almost unlimited number of terms, interconnected by logical operations "AND", "OR" or "AND-NOT". Each search term refers to one or another user-selected metadata element (field) of the object. The field, depending on the information it contains, is assigned to a certain type. Depending on what type the field belongs to, the user is given the opportunity to set certain conditions for selecting its content.

The system provides the following types of fields and their corresponding search terms:

- Text field. This type includes fields that contain relatively short text information. For this type, it is possible to specify text fragments with the qualification "contains", "starts with" and "equal" (by default, the condition "contains" is used).

- Full text field. This type includes fields that contain significant textual information, for example, a biography of a scientist. For this type of field, it is possible to specify word forms that must be present in the search field. PostgreSQL full-text search tools are used to organize the search.
- Numeric field. For fields of this type, it is possible to specify the condition “greater than”, “equal” or “less” and the numeric value with which you want to compare. In some cases, the values of numeric fields in the database may not be determined (person's year of birth, book publication year, etc.). The system allows entering into these fields non-numeric values for example “mid-17th century”. When specifying conditions for numeric fields, records with undefined values will never be included in the resulting output. To be able to ensure the completeness of the search, the option “Add documents with an unspecified value” has been added for numeric fields. When this option is specified, both documents whose fields meet the specified condition and documents whose field value is not numeric are selected.
- Short codifier. This type includes fields whose value is limited to a certain list, for example, the language of the publication or its type. To set conditions for a field of this type, the user is given the entire corresponding codifier and given the opportunity to mark the desired values.
- Hierarchical codifier. Unlike a short codifier, this type can have a significant volume and hierarchical structure. When choosing elements from this codifier, the user is given the opportunity to preliminary search both by the names of the headings and by their code. For codifier elements that have subordinate headings, it is possible to specify whether to include subordinate headings in the request or to limit itself to documents whose encoding contains only this element. In the system, an example of such a codifier is the GRNTI codifier.
- Date. This type includes fields whose value is a date, for example, the date the library was received. To specify conditions for a field of this type, the user is provided with the opportunity to specify the required date and the condition “greater than”, “equal to” or “less than”.

For queries with a large volume of results obtained, the possibility of pagination is provided. In this case, the search in the database is carried out once and the result of the query is remembered. This allows, when scrolling through the pages, not to perform a second search, but to promptly display the data of the next page.

The capabilities listed above allow you to form a query for fields related to a specific metadata element. An exception is the ability to search for publications by the author's last name. Those the formulation of conditions for publication allows you to set a condition for the presence of a person associated with it with the required surname.

In the new system, an attempt is made to provide the user with a universal search option, in which it is possible to set conditions not only for a certain type of metadata, but also for the metadata associated with this type, and, in turn, for the metadata associated with associated metadata, etc.

The solution to this problem required the development of both an intuitive search query builder and a strategy for generating a query to the database, which, on the one

hand, would provide simple algorithmization and, on the other hand, sufficient efficiency of the generated query. Both tasks are solved in the new system. In Fig. 3 shows a screenshot with an example of a request for finding Euler's coauthors.

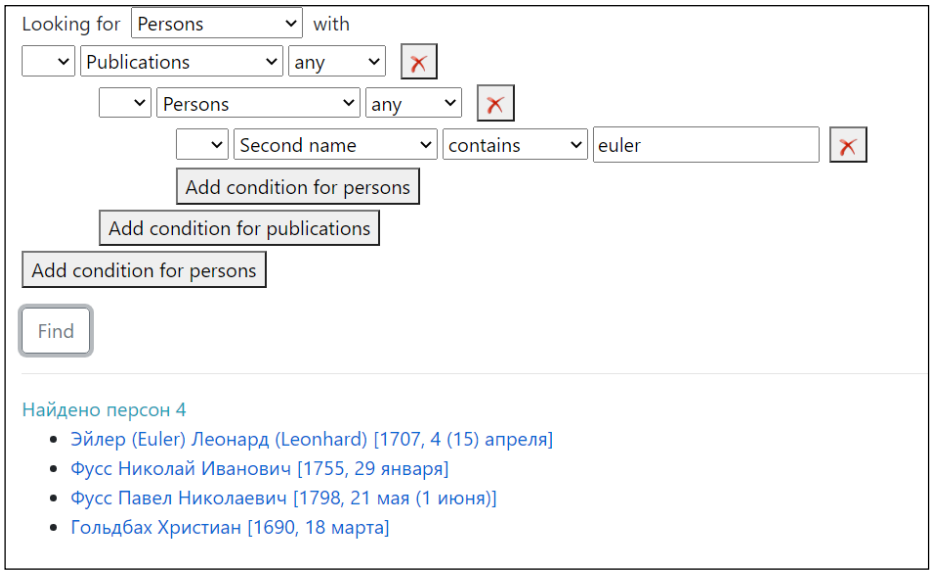

Fig. 3. Search result for persons related with publications to Euler.

The new version of the system is currently in trial operation. The current version of the Electronic Library “Scientific Heritage of Russia” continues to function and is available at http://e-heritage.ru. To switch to the new system, a separate program for transferring data from the databases of the old system to the database of the new system has been implemented. After revision and final testing of the new version, it will replace the previous one without any interruption in work.

The approaches worked out during the implementation of the system can be used as the basis for the basic software to support various components of a single digital space of scientific knowledge [6, 7].

The work was carried out at the Joint Supercomputer Center of Russian Academy of Sciences in the framework of state assignment No. 0580–2021–0014.

References

1. Kalenov, N.E., Savin, G.I., Sotnikov, A.N. Tehnologiya sozdaniya elektronnoy biblioteki “Nauchnoye naslediye Rossii”. Nauchnaya kniga 1-4. S. 170–173 (2007).
2. Kalenov, N.E., Savin, G.I., Serebryakov, V.A., Sotnikov, A.N.: Printsipy postroyeniya i formirovaniya elektronnoy biblioteki “Nauchnoye naslediye Rossii”. Programmnyye Produkty, Sistemy i Algoritmy 4 (100). S. 30–40 (2012).
3. Pogorelko, K.P.: Dinamika ispol'zovaniya elektronnoy biblioteki “Nauchnoye naslediye Rossii”. In: Informatsionnoye obespecheniye nauki: novyie tekhnologii: Sbornik nauchnykh trudov. Moskva: BEN RAN. S. 192–200 (2017).

4. Pogorelko, K.P.: Novaya systema prezentatsii elektronnykh knig v systeme “Nauchnoye naslediye Rossii”. In: Informatsionnoye obespecheniye nauki: novyye tekhnologii: Sbornik nauchnykh trudov. Moskva: BEN RAN. S. 32–35 (2013).
5. Kalenov, N.E., Kirillov, S.A., Sobolevskaya, I.N., Sotnikov, A.N.: Sovremennoye sostoyaniye elektronnoy biblioteki “Nauchnoye naslediye Rossii”. Trudy NIISI RAN. Matematicheskoye i komp'yuternoye modelirovaniye slozhnykh sistem: teoreticheskiye i prikladnyye aspekty 8 (6). S. 166–169 (2018).
6. Antopol’skij, A.B., Kalenov, N.E., Serebryakov, V.A., Sotnikov, A.N.: O edinom tsifrovom prostranstve nauchnykh znaniy. Vestnik Rossiyskoy Akademii Nauk 89 (7). S. 728–735 (2019).
7. Antopol'skij, A.B. i dr.: Printsipy postroyeniya i struktura yedinogo tsifrovogo prostranstva nauchnykh znaniy. Nauchno-tekhnicheskaya Informatsiya. Seriya 1: Organizatsiya i Metodika Informatsionnoy Raboty (4). S. 9–17 (2020).