Развитие программного обеспечения электронной библиотеки «Научное наследие России»

К. П. Погорелко

МСЦ РАН

Аннотация. Структура данных и программное обеспечение электронной библиотеки «Научное наследие России» создавалось в 2007 г. и в настоящее время не отвечает потребностям системы. В статье описаны решения, принятые при реализации нового варианта программного обеспечения. Эти решения затрагивают как организацию структуры базы данных, так и протоколы взаимодействия с системой. Особое внимание уделяется разработке средств поиска информации.

Ключевые слова: электронная библиотека, цифровая библиотека, программное обеспечение, поиск информации, asp.net core.

Development of the Software for the Electronic Library «Scientific Heritage of Russia»

K. P. Pogorelko

JSCC RAS

Abstract. The data structure and software of the Scientific Heritage of Russia electronic library was created in 2007 and currently does not meet the needs of the system. The article describes the decisions made when implementing a new version of the software. These decisions affect both the organization of the database structure and the protocols for interacting with the system. Particular attention is paid to the development of information retrieval tools.

Keywords: electronic library, digital library, software, information search, asp.net core.

Электронная библиотека (ЭБ) «Научное наследие России» начала создаваться в 2007 г. [1] и функционирует в полном объеме более 10 лет [2]. Библиотека пользуется постоянным спросом [3]. Распределение по годам количества запросов к электронным публикациям, включенным в ЭБ, приведено на *рис. 1*.

Программное обеспечение, поддерживающее функционирование библиотеки, несколько раз модифицировалось. Так, в 2013 г. представление полнотекстовых публикаций было выделено в отдельную подсистему [4]. Современное состояние библиотеки приводится в [5].

Рис. 1 Количество запросов на электронные публикации

В настоящее время назрела необходимость переработки программного обеспечения. Это обусловлено как возросшими требованиями к поисковым возможностям системы и ее быстродействию, так и к более полному удовлетворению требований, выдвигаемых технологическими процессами подготовки информации.

Поскольку средства разработки, на которых базируется существующая система, устарели, было принято решение о коренной переработке как программного обеспечения,

поддерживающего работу системы, так и ее внутренней структуры и интерфейсов.

Для реализации новой версии программного обеспечения была выбрана современная платформа asp.net core. Основной особенностью данной платформы является универсальность приложений, разработанных на ее базе. Эти приложения могут функционировать как на операционных системах Microsoft, так и на операционных системах семейства Unix. Кроме того, при развертывании веб-приложений, разработанных на базе этой платформы, они оформляются как исполняемый файл, что обеспечивает их высокую эффективность. Для хранения данных был выбран SQL сервер PostgreSQL, который имеет реализации для разных платформ и обеспечивает достаточно эффективное манипулирование данными.

Разрабатываемая система объединена в единый проект, состоящий из следующих частей: подсистема подготовки метаинформации по различным объектам, подсистема подготовки электронных публикаций и подсистема поиска и визуализации информации. Такое решение позволяет учитывать возможные изменения состава и структуры данных сразу во всем проекте. Кроме того, в зависимости от производительности серверов и нагрузки на систему, появляется возможность разворачивания системы как на одном сервере, так и разнесения составляющих проекта по разным серверам.

В предыдущей версии системы информация располагалась в нескольких базах данных, между которыми были организованы протоколы обмена. Помимо задержек с публикацией и увеличением вероятностей сбоя, это создавало определенные трудности в организации перекрестных ссылок между объектами, которые были размещены в разных базах. В новой версии системы все данные содержатся в единой базе, что обеспечивает единообразие их обработки, облегчает манипулирование данными и обеспечивает отсутствие задержек при их публикации.

Для обеспечения контроля качества вводимых данных принята двухуровневая система полномочий. На первом

уровне оператор, имеющий полномочия на ввод данных, осуществляет их ввод. По окончании ввода оператор устанавливает признак «ввод окончен». Выпускающий редактор, просматривая документы с признаком окончания ввода, проверяет введенные данные и, если данные введены с необходимым качеством, устанавливает признак «опубликовано». После того как документ получает этот статус, он становится доступным для поиска и показа в библиотеке. Оператор, осуществивший ввод документа, получившего статус «опубликовано», теряет право коррекции данного документа, если он не имеет прав выпускающего редактора.

В системе реализованы средства для поддержки процедур выявления и обработки ошибочных данных. Ошибки могут выявляться как в ручном режиме при просмотре данных, так и в автоматическом режиме, например, при проверке

корректности внешних ссылок. Все выявленные ошибки приписываются в виде списка к тому объекту, к которому они относятся. Оператор имеет возможность просматривать выявленные ошибки и, после их исправления, присваивать документу соответствующий статус.

В структуре базы данных предусмотрена возможность снабжать связи между метаданными различных объектов дополнительными атрибутами, конкретизирующими данную связь. Например, для связи объектов «персона» — «публикация» можно указать значение атрибута «автор», «редактор», «о нем» и т. д.

В новой версии системы существенно расширены ее поисковые возможности. Поисковый модуль системы позволяет обрабатывать запросы, состоящие из практически неограниченного количества термов, связанных между собой логическими операциями «И», «ИЛИ» или «И-НЕ». Каждый поисковый терм относится к тому или иному выбранному пользователем элементу метаданных (полно) объекта. Поле в зависимости от информации, которую оно содержит, относится к некоторому типу. В зависимости от того, к какому типу относится поле, пользователю предоставляется возможность задать те или иные условия для отбора его содержимого.

В системе предусмотрены следующие типы полей и соответствующие им поисковые условия:

1. Текстовое поле. К этому типу относятся поля, содержащие относительно короткую текстовую информацию. Для этого типа имеется возможность задать фрагменты текста с уточнением «содержится», «начинается с» и «равно» (по умолчанию используется условие «содержится»).

2. Полнотекстовое поле. К этому типу относятся поля, содержащие значительную по объему текстовую информацию, например, биографию ученого. Для такого типа полей имеется возможность задать словоформы, которые должны присутствовать в поисковом поле. Для организации поиска применяются средства полнотекстового поиска PostgreSQL.

3. Числовое поле. Для полей такого типа имеется возможность задать условие «больше», «равно» или «меньше» и числовое значение, с которым необходимо производить сравнение. В ряде случаев значения числовых полей в базе данных могут быть не определены (год рождения персоны, год издания книги и др.). Система допускает ввод в эти поля нечисловых значений типа «середина 17-го века» и т. п. При задании условий на числовые поля записи с неопределенными значениями никогда не попадут в результирующую выдачу. Чтобы иметь возможность обеспечить полноту поиска, для числовых полей добавлена опция «Добавить документы с неустановленным значением». При задании этой опции выбираются как документы, поля которых отвечают заданному условию, так и документы, у которых значение поля не является числовым.

4. Кодификатор. К этому типу относятся поля, значение которых ограничено определенным перечнем, например, рубрика ГРНТИ или тип публикации. Для задания условий на поле этого типа пользователю выдается соответствующий кодификатор и предоставляется возможность отметить нужные значения.

5. Дата. К этому типу относятся поля, значение которых является датой, например, дата поступления в библиотеку. Для задания условий на поле этого типа поль-

зователю обеспечивается возможность задать необходимую дату и условие «больше», «равно» или «меньше».

Для запросов с большим объемом полученных результатов предусмотрена возможность постраничной выдачи. При этом поиск в базе осуществляется один раз и результат выполнения запроса запоминается. Это позволяет, при пролистывании страниц, не осуществлять повторный поиск, а оперативно выдавать данные очередной страницы.

Перечисленные выше возможности позволяют формировать запрос к полям, относящимся к определенному элементу метаданных. Исключением является возможность поиска публикаций по фамилии автора. В новой системе реализована попытка предоставления пользователю возможности универсального поиска, при которой можно задавать условия не только к определенному типу метаданных, но и к метаданным, связанным с этим типом, и, в свою очередь, к метаданным, связанным со связанными метаданными и т. д. Решение этой задачи потребовало разработки как интуитивно понятного построителя поискового запроса, так и стратегии формирования запроса к базе данных которая, с одной стороны, обеспечивала бы простую алгоритмизацию и, с другой стороны, достаточную эффективность сформированного запроса. Обе задачи в новой системе решены. На *рис. 2* приведен скриншот с примером запроса на поиск соавторов Эйлера.

Ищем у которых

Найдено персон 4

- Штриттер Иван Михайлович [1740, 10 (21) октября]
- Фусс Павел Николаевич [1798, 21 мая (1 июня)]
- Гольдбах Христиан [1690, 18 марта]
- Фусс Николай Иванович [1755, 29 января]

Рис. 2. Пример выполнения запроса к универсальной системе поиска

Новая версия системы в настоящее время находится в стадии опытной эксплуатации. Текущая версия ЭБ «Научное наследие России» продолжает функционировать и доступна по адресу <http://e-heritage.ru>. Для перехода на новую систему реализована отдельная программа переноса данных из баз старой системы в базу данных новой системы. После доработки и окончательного тестирования новой версии она заменит предыдущую без каких-либо перерывов в работе.

Обрабатываемые при реализации системы подходы могут быть положены в основу базового программного обеспечения для поддержки различных компонентов единого цифрового пространства научных знаний [6, 7].

Работа выполнена в МСЦ РАН в рамках государственного задания № 0580–2021–0014.

Литература

1. Каленов Н. Е., Савин Г. И., Сотников А. Н. Технология создания электронной библиотеки «Научное наследие России» // Научная книга. 2007. Т. 1–4. С. 170–173.
2. Каленов Н. Е., Савин Г. И., Серебряков В. А., Сотников А. Н. Принципы построения и формирования электронной библиотеки «Научное наследие России» // Программные продукты, системы и алгоритмы. 2012. Т. 4. № 100. С. 30–40.
3. Погорелко К. П. Динамика использования электронной библиотеки «Научное наследие России» // Информационное обеспечение науки: новые технологии: Сборник научных трудов. М.: БЕН РАН, 2017. С. 192–200.
4. Погорелко К. П. Новая система презентации электронных книг в системе «Научное наследие России» // Информационное обеспечение науки: новые технологии. Сборник научных трудов / Каленов Н. Е. (ред.). М.: БЕН РАН, 2013. С. 32–35.
5. Каленов Н. Е., Кириллов С. А., Соболевская И. Н., Сотников А. Н. Современное состояние электронной библиотеки «Научное наследие России» // Труды НИИСИ РАН. Математическое и компьютерное моделирование сложных систем: теоретические и прикладные аспекты. 2018. Т. 8. № 6. С. 166–169.
6. Антопольский А. Б., Каленов Н. Е., Серебряков В. А., Сотников А. Н. О Едином цифровом пространстве научных знаний // Вестник Российской академии наук. 2019. Т. 89. № 7. С. 728–735.
7. Антопольский А. Б. и др. Принципы построения и структура Единого цифрового пространства научных знаний (ЕЦПЗ) // Научно-техническая информация. Серия 1: Организация и методика информационной работы. 2020. № 4. С. 9–17.

References

1. Kalenov N. E., Savin G. I., Sotnikov A. N. Technology of creating an electronic library «Scientific heritage of Russia» // Nauchnaya kniga. 2007. T. 1–4. P. 170–173.
2. Kalenov N. E., Savin G. I., Serebryakov V. A., Sotnikov A. N. Principles of construction and formation of the electronic library «Scientific heritage of Russia» // Software products, systems and algorithms. 2012. T. 4. № 100. P. 30–40.
3. Pogorelko K. P. Dynamics of the use of the electronic library «Scientific heritage of Russia» // Information support of science: new technologies: Collection of scientific papers. M.: LNS RAS, 2017. P. 192–200.
4. Pogorelko K. P. New system of presentation of e-books in the system «Scientific heritage of Russia» // Information support of science: new

technologies. Collection of scientific papers / Kalenov N. E. (ed.). M.: LNS RAS, 2013. P. 32–35.

5. Kalenov N. E., Kirillov S. A., Sobolevskaya I. N., Sotnikov A. N. The current state of the electronic library «Scientific heritage of Russia» // Proceedings of NIISI RAS. Mathematical and computer modeling of complex systems: theoretical and applied aspects. 2018. T. 8. № 6. P. 166–169.

6. Antopol'skij A. B., Kalenov N. E., Serebryakov V. A., Sotnikov A. N. On the Unified Digital Space of Scientific Knowledge // Bulletin of the Russian Academy of Sciences. 2019. T. 89. № 7. P. 728–735.

7. Antopol'skij A. B. et al. Principles of construction and structure of the Unified digital space of scientific knowledge (UDSSK) // Scientific and technical information. Series 1: Organization and methodology of information work. 2020. № 4. P. 9–17.