Информационная система архивных документов ВАК – результаты за 2018 год

VAK information system for archived documents – 2018 results

А. И. Бродовский, Е. М. Зайцева, Ю. И. Заславский, Б. И. Маршак
Государственная публичная научно-техническая библиотека России, Москва, Россия

Alexander Brodovsky, Ekaterina Zaitseva, Yury Zaslavsky and Boris Marshak
Russian National Public Library for Science and Technology, Moscow, Russia

Описывается комплекс работ по разработке и ведению информационной системы архивных документов ВАК. Отражаются результаты анализа архивных материалов, принципы отбора материалов для оцифровки, процесс верификации документов. Дается описание технологии оцифровки архивных материалов, самой информационной системы архивных документов ВАК, ее структуры и поисковых возможностей. Рассмотрены вопросы технологической обработки и хранения файлов отсканированных документов. Представлены результаты работ по проекту за 2016–2018 гг. с акцентом на результаты работ в 2018 г.
Ключевые слова: информационная система архивных документов ВАК, оцифровка, поиск, хранение, верификация данных

The authors describe the process of design and maintenance of VAK (The Higher Attestation Commission) information system for archived documents. They discuss the results of archival materials analysis, selection principles of digitization, and document verification process. The digitization process, VAK information system, its structure and functionalities are characterized. The authors also examine the issues of technological processing and storage of scanned documents. The Project results for 2016–2018 are presented with the main focus on the year 2018.
Keywords: VAK information system for archival documents, Higher Attestation Commission (VAK), digitization, search, storage, data verification.

Информационная система архива является наиболее перспективной разновидностью реализации архива среди действующих структурных типов электронных архивов [1]. Она позволяет создавать, хранить документы, управлять ими и осуществлять по ним поиск. Именно путь создания информационной системы архива был выбран при реализации проекта формирования информационных ресурсов архивных документов Высшей аттестационной комиссии (ВАК) при Минобрнауки России.

С 2016 г. в ГПНТБ России ведется активная работа по разработке, развитию и обеспечению функционирования информационной системы архивных документов ВАК, включая оцифровку архивных документов, обеспечение поиска по архиву, анализ и верификацию архивных материалов. Базовой целью проекта является обеспечение функционирования и развития государственной системы научной аттестации, что предполагает формирование информационных ресурсов архивных документов ВАК при Минобрнауки России, обеспечение доступа к архивным документам работников Минобрнауки России и экспертных советов ВАК, повышение эффективности и качества работы подразделений Департамента аттестации научных и научно-педагогических работников Минобрнауки России и экспертной работы за счет использования в работе данных оцифрованного архива документов.

Особенности создаваемой информационной системы обусловлены прежде всего специфическими типами документов, хранящимися в архиве ВАК, а также тем, что основными задачами при проектировании системы были реализация принципа простоты и удобства интерфейса, принципа оперативности поиска и просмотра необходимых документов, а также обеспечение возможности копирования документов и подготовки на основе архивных материалов различного вида справок и отчетов. Создаваемый электронный архив призван выполнять следующие основные функции: хранение электронных документов, эффективный поиск информации, оперативный доступ к документам, управление документами (печать, выгрузка). При реализации проекта разработчики

системы старались учесть опыт создания электронных архивов другого вида и различных подходов к их организации и хранению [2, 3, 4, 5]. В последние годы по вопросам создания электронных архивов сформировалась определенная нормативная и методологическая база [6, 7, 8, 9, 10], на которую также опирались разработчики системы.

Основываясь на имеющемся опыте создания электронных архивов, на сложившейся нормативной и методологической базе, в ходе реализации проекта специалисты ГПНТБ России разработали технологию обработки архивных материалов и программное и информационное обеспечение системы с учетом особенностей архивных документов ВАК и требований Минобрнауки России.

По проекту выполняется следующий комплекс работ:

1) анализ и верификация материалов;

2) оцифровка архивных материалов;

3) разработка программного и информационного обеспечения информационной системы архивных документов;

4) технологическая обработка отсканированных документов;

5) обеспечение хранения оцифрованных материалов;

6) вспомогательные работы.

1. Анализ и верификация архивных материалов

Содержательная обработка архивных материалов включает следующие работы:

1) Анализ состояния архива.

Были просмотрены и проанализированы различные архивные материалы ВАК, включая картотеку ВАК, приказы Минобрнауки России, решения Президиума ВАК, заключения Президиума ВАК, рекомендации Президиума ВАК, протоколы заседаний Президиума ВАК и т.д.

2) Разработка принципов отбора и отбор архивных материалов для оцифровки.

В качестве основных критериев отбора документов для первичного наполнения базы данных информационной системы архивных документов ВАК были выбраны следующие:

а) отражение в архивных документах основных сведений о лице, имеющем ученую степень/звание;

б) востребованность данных, содержащихся в архивных документах, в деятельности работников Минобрнауки России и экспертных советов ВАК;

в) степень сохранности архивных документов;

г) качество оформления архивных документов (наличие/отсутствие рукописного текста).

Среди большого массива архивных документов ВАК в наибольшей степени перечисленным критериям удовлетворяют приказы Минобрнауки России и решения Президиума ВАК, которые и были выбраны в качестве базовых материалов для первичного наполнения информационной системы архивных документов ВАК в 2016 г. В 2017 г. спектр охвата архивных документов был расширен. С учетом рекомендаций сотрудников Департамента аттестации научных и научно-педагогических работников Минобрнауки России были определены типы основных, значимых и контрольных типов архивных документов ВАК, которые обрабатывались в полном объеме в 2017–2018 гг.

С учетом выделенных типов документов при выполнении работ по проекту в 2018 г. была осуществлена модернизация базы данных и поискового функционала информационной системы архивных документов ВАК с реализацией поиска по типу документа. Также была выявлена необходимость реализации поиска по признаку года составления документа, отраженного в названии папки документов. Год составления документа, зафиксированный в тексте самого документа, не является надежным поисковым элементом, поскольку может быть проставлен на документе от руки. Отдельно организованный поиск по году составления документа, зафиксированному в базе данных, снимает эту проблему.

3) Верификация архивных материалов.

Полученные из Минобрнауки России документы перед сканированием проверяются на дублетность, лакуны, правильность порядка расположения, последовательную простановку дат и номеров документов; комплекты документов проверяются на полноту подборки; удаляются лишние страницы (копии, черновые страницы); в отдельных случаях проводится перекомпоновка

материалов по согласованию с Департаментом аттестации научных и научно-педагогических работников Минобрнауки России.

2. Оцифровка архивных материалов

В ходе выполнения проекта была разработана технология процесса оцифровки архивных документов. Предварительно были сформулированы требования, предъявляемые к файлам электронного архива, определены формат и технологические параметры файлов, содержащих графические образы.

Технологические параметры файлов, содержащих графические образы, определялись с учетом общепринятых стандартов и требований сканирования, читабельности получаемых изображений для пользователя, а также требований информационной системы архивных документов. В качестве основных параметров были установлены следующие: тип файлов – файлы JPEG; размер файлов – не более 1Мбайт; оптическое разрешение – 300x300 dpi; глубина цвета – 8 бит.

Основные технологические этапы оцифровки определяются следующим образом:

1) Первичный осмотр и структурирование исходных материалов.

Перед сканированием проводится осмотр и экспертиза архивных материалов с оценкой качества их оформления. Выявляются документы, имеющие различного рода дефекты и искажения и требующие применения не поточного, а отдельного постраничного сканирования.

2) Сканирование документов.

Сканирование проводится на специализированных сканерах при специальном режиме, который практически исключает инфракрасное и ультрафиолетовое воздействие на оригинал, ограничивает до минимального световое воздействие, что исключает порчу бумажных оригиналов. В процессе работы используются два сканера: высокопроизводительный поточный сканер фирмы «Kodak», обеспечивающий качественную и оперативную оцифровку документов и планетарный сканер фирмы «Zeutschel», применяемый для бережного постраничного сканирования литературы. В результате сканирования получаются массивы файлов типа JPEG.

3) Обработка и проверка полученных образов.

Осуществляется полистная проверка оцифрованных документов, корректировка и чистка файлов с устранением имеющихся дефектов, отклонений, затемнений и т.д.

4) Структурирование оцифрованного массива.

В процессе работы проводится формирование папок с именами, включающими тип содержащихся в папке документов и даты их создания, нумерация файлов и распределение их по папкам в соответствии с томами архивных материалов.

5) Выходной контроль качества массивов графических образов.

На выходе проводится общий контроль качества результатов оцифровки, включающий комплекс проверок различного вида: проверка оформления, наименования и нумерации папок и файлов; сравнение количества образов с количеством страниц документов-оригиналов; проверка на отсутствие пропусков и дублей; выравнивание образов по размеру; проверка качества графических образов; контроль расфокуса («размытого» изображения); отсутствие загибов страниц; контроль обрезки текста; наличие полей по краям; определение наклона текста; точная ориентация по тексту (доворот); исправление геометрических искажений текста; удаление затемнений и теней; ликвидация пятен, мусора, посторонних объектов.

3. Разработка программного и информационного обеспечения информационной системы архивных документов ВАК

Разработка информационной системы включала следующие основные этапы:

1) разработка информационной системы архивных документов ВАК на основе системы автоматизации библиотек (САБ) ИРБИС64, включая:

- разработку поискового интерфейса, обеспечивающего просмотр папок архивных материалов, поиск по папкам архивных материалов и по любому текстовому элементу документа ВАК (ФИО, ученой степени/званию, названию научного учреждения и др.);
- разработку форматов;

2) установка информационной системы в Минобрнауки России;

3) модернизация структуры базы данных информационной системы и поискового функционала.

Основой программного обеспечения информационной системы «Архивные документы ВАК» (в дальнейшем – Система) является САБ ИРБИС64. При создании Системы использовались следующие модули и технологии САБ ИРБИС64: сервер баз данных ИРБИС64; технология создания базы данных имидж-каталога; модуль ИРБИС-Навигатор.

Базой данных Системы является имидж-каталог. Имидж-каталог представляет собой полнотекстовую базу данных, созданную на основе распознанных скан-образов текстовых документов. Для Системы документом является одна страница архива ВАК.

Процесс создания базы данных Системы состоит из двух этапов: постраничное сканирование архивных документов ВАК; формирование базы данных имидж-каталога на основе сканированных образов страниц, включающее процесс автоматического распознавания их текстов.

В результате сканирования образ каждой страницы сохраняется в виде графического файла в формате JPEG, а образы всех страниц из одной архивной папки (книги) помещаются в папку (директорию на диске) с именем, соответствующим содержанию архивной папки.

Процесс формирования базы данных имидж-каталога представляет собой полностью пакетную обработку. Для этого используются специальные средства САБ ИРБИС64, включающие функцию автоматического распознавания текста.

Пользователю для работы с Системой предлагается ИРБИС-Навигатор. Данный модуль представляет собой клиентское приложение, предназначенное для выполнения произвольных операций с базами данных САБ ИРБИС64 на основе интерфейсов, программируемых с помощью форматов САБ ИРБИС64. Форматами САБ ИРБИС64 называются сценарии представления данных, состоящие из конструкций языка форматирования (языка манипулирования данными) САБ ИРБИС64 и HTML-тэгов. В рамках описываемой работы для ИРБИС-Навигатора был разработан набор форматов, реализующих весь функционал Системы, адресованный пользователю.

Принцип работы и пользовательский интерфейс ИРБИС-Навигатора аналогичен Web-браузеру. Основной поисковый экран Системы представлен на рис. 1.

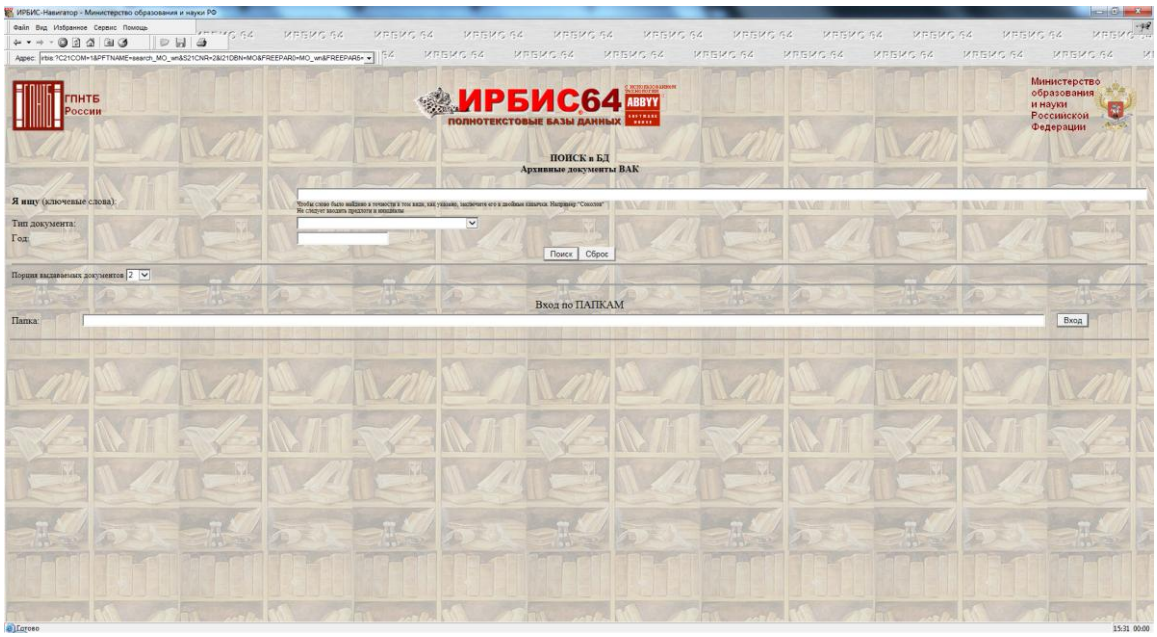

Рис. 1 – Основной поисковый экран информационной системы «Архивные документы ВАК»

Интерфейс предлагает для поиска две основные возможности:

1) Полнотекстовый поиск страниц архива по словам – с помощью редактируемой строки «Я ищу». Данный вид поиска предполагается использовать как основной.

2) Поиск путем последовательного просмотра архивных страниц в определенной папке – с помощью редактируемой строки «Папка».

В 2018 г. также реализованы возможности задания при поиске типа документа и года составления документа.

Запрос для полнотекстового поиска формулируется на естественном языке. В качестве терминов запроса можно использовать любые слова, содержащиеся в архивных документах: ФИО, ученая степень/звание, название научного учреждения и т.д.

Результат полнотекстового поиска (после нажатия кнопки «Поиск») представляется в виде последовательности найденных страниц архива. Найденные страницы располагаются в порядке убывания их релевантности (соответствия запросу). Релевантность определяется на основе оригинального критерия, который учитывает количество и контекстную близость слов запроса, найденных на странице архива. Термины (слова), найденные на страницах архива, маркируются цветом.

В случае полнотекстового поиска для каждого найденного документа (страницы архива) указывается название папки, в которой он находится, и, кроме того, даются ссылки, позволяющие определить контекст данной страницы в папке, а именно – возможность перейти к первой, предыдущей, следующей странице архива.

Для выполнения поиска по папкам необходимо в соответствующий элемент основного поискового интерфейса ввести полное название папки или его префиксную часть и нажать кнопку «Вход». В результате Система выдает алфавитный список ссылок на папки, чьи названия соответствуют введенным данным. Для конкретной папки предлагается порционный просмотр ее страниц в порядке их следования.

Для любого выдаваемого документа (страницы архива) предлагается список функциональных возможностей: сохранить документ, распечатать документ, отправить документ по почте и др.

4. Технологическая обработка отсканированных документов

Технологическая обработка отсканированных документов включает следующие основные этапы:

1) "Скрытое" распознавание текста, включающее: разбивка текста на предложения; извлечение слов; удаление стоп-слов (неинформативных слов).

2) Формирование базы данных информационной системы архивных документов ВАК, включающее: формирование записей с координатами слов; формирование словаря. База данных информационной системы архивных документов ВАК формируется как стандартная база данных САБ ИРБИС64.

3) Пополнение базы данных информационной системы архивных документов ВАК. Пополнение базы данных происходит по мере оцифровки и технологической обработки отсканированных документов.

5. Обеспечение хранения оцифрованных материалов

В ходе выполняемых работ соблюдается определенный порядок размещения, передачи и хранения файлов информационной системы архивных документов: сканирование документов архива и сохранение файлов JPEG в рабочей папке; проверка, корректировка, очистка и структурирование файлов в рабочей папке; копирование готовых файлов в папку результатов; проверка файлов в папке результатов; передача папки результатов на технологическую обработку в информационной системе, а также на хранение на сервере и создание резервной копии; актуализация базы данных архивных документов в информационной системе и передача базы данных архивных документов на хранение; передача базы данных и отсканированных архивных документов в Минобрнауки России.

6. Вспомогательные работы

Вспомогательные работы по проекту включают следующие операции: расшивка томов архивных материалов; обрезка листов; механическая очистка страниц от нитей и клея; переплетные работы.

В ходе выполнения проекта получены следующие основные результаты:

1) Создана информационная система архивных документов ВАК.

2) Разработана технология оцифровки архивных документов.

3) В 2018 г. проведена модернизации базы данных и поискового функционала информационной системы архивных документов ВАК с реализацией поиска по типу документа и году его составления.

4) В 2018 г. проведена верификация базовых архивных документов (приказов Минобрнауки России и решений Президиума ВАК) с 2017 г. до 2000 г.

5) В ходе реализации проекта за 2016-2018 гг. оцифровано и обработано 1011 томов архивных документов ВАК (2016 г. – 448 томов, 2017 г. – 265 томов, 2018 г. – 298 томов).

Загруженные в информационную систему архивные данные могут эффективно использоваться в деятельности Минобрнауки России и экспертных советов ВАК для получения любой справочной информации, касающейся присуждения ученых степеней и присвоения ученых званий, а также получения наукометрических показателей деятельности научных учреждений.

Список источников

1. Рындин А. А. Архив без пыльных полок или способы организации архива документов предприятия [Электронный ресурс] // ECM-Journal: журнал о системах электронного документооборота (СЭД). – Режим доступа: https://ecm-journal.ru/card.aspx?ContentID=1912029. – Загл. с экрана.
2. Юмашева Ю.Ю. Архивы электронных документов: проблемы и возможные решения // Власть. – 2015. – № 3. – С 61-66.
3. Залаев Г. З., Каленов Н. Е., Цветкова В. А. Оцифровка документов в научных архивах и библиотеках: вопросы и ответы // НТИ. Сер.1. – 2016. – № 2. – С. 14-21.
4. Тихонов В.И. Архивное хранение электронных документов: проблемы и решения [Электронный ресурс] // Делопроизводство и документооборот на предприятии, февраль 2006. – Режим доступа: http://www.delo-press/articles.php?n=5150. – Загл. с экрана.
5. Евстигнеева Г. А. Качество оцифровки – проблемы и решения // Современная библиотека. – 2012. – № 5. – С. 58-61.
6. ГОСТ Р ИСО 30300-2015. Система стандартов по информации, библиотечному и издательскому делу. Информация и документация. Системы управления документами. Основные положения и словарь. – Введ. 2016-07-01. – М.: Стандартинформ, 2016. – 14 с.
7. ГОСТ Р ИСО 15489-1-2007. Система стандартов по информации, библиотечному и издательскому делу. Управление документами. Общие требования. – Введ. 2007-07-01. – М.: Стандартинформ, 2007. – 20 с.
8. Правила организации хранения, комплектования, учета и использования документов Архивного фонда Российской Федерации и других архивных документов в органах государственной власти, органах местного самоуправления и организациях (утверждены приказом Министерства культуры РФ от 31 марта 2015 г. № 526) // Электронный фонд правовой и нормативно-технической документации. – Режим доступа: http://docs.cntd.ru/document/420266293. – Загл. с экрана.
9. Методические рекомендации по электронному копированию архивных документов и управлению полученным информационным массивом / Ю. Ю. Юмашева. – М.: ВНИИДАД, 2012. – 125 с.
10. Методика контроля качества сканирования бумажных документов: методическое пособие и техническое руководство / С. М. Тимиргалиев, Н. И. Черновалова, О. В. Баркова, Е. В. Ларкин, В. В. Котов, С. Н. Клещарь, Ю. И. Заславский. – М.: ДиМи-Центр, 2012. – 53 с.