УДК 004.891.3 DOI: 10.18698/2541-8009-2021-12-754

РАЗРАБОТКА СИСТЕМЫ ДИНАМИЧЕСКОГО АРХИВИРОВАНИЯ

Алия Мухаммад eng.muhammadaliah@gmail.com
SPIN-код: 4669-7154

МГТУ им. Н.Э. Баумана, Москва, Российская Федераци

Аннотация

Управление документами является важной частью работы организаций любого уровня в различных сферах деятельности и включает в себя управление контентом, цифровыми активами, ксерокопированием и документооборотом. В статье описан процесс разработки системы динамического архивирования, позволяющей создавать категории документов и их атрибуты, проводить оптическое распознавание символов и, как следствие, повышающей эффективность управления документами. С ее помощью компании смогут самостоятельно выстраивать оптимальную модель управления документами независимо от их атрибутов, содержимого, времени и места хранения. Цель достигается благодаря возможности поиска документов по указанным для них тегам или по их содержимому, обеспечения оптического распознавания символов. Кроме того, разработана модель разграничения доступа пользователей, а также обеспечена возможность добавления текста в файлы изображений.

Ключевые слова

Система динамического архивирования, оптическое распознавание символов, сноски, документы, редактирование фотографий, диалоговая модель, база данных, ролевой контроль доступа, компьютерные системы

Поступила в редакцию 13.05.2021

Введение. Архивирование данных означает их долговременное хранение с возможностью многократного повторного использования. Данные, как правило, собираются и хранятся для выполнения операций в реальном времени и управления накопленной информацией [1-2].

Система управления документами — это компьютерная система, используемая для отслеживания и хранения электронных документов или копий бумажных документов, в том числе в нескольких версиях. Система управления документами управляет корпоративным контентом, цифровыми активами, ксерокопированием и документооборотом [3].

Проблема эффективного управления документами в современных организациях связана с несколькими аспектами:

1) обеспечение оптимального баланса между количеством и качеством документации;

2) трудности поиска и извлечения документов (как следствие, лишние затраты времени);

3) вопросы хранения бумажных документов, которые занимают большую площадь ограниченного пространства, которое необходимо грамотно использовать;

4) проблемы обеспечения защиты и безопасности (нарушение конфиденциальности, целостности, доступности).

Перечисленные проблемы актуальны для организации любого уровня в различных сферах деятельности. Их решение возможно с помощью разработки и внедрения системы динамического архивирования.

Например, в больнице имеется большое количество неструктурированной информации о пациентах: архивные данные, персональные данные, история болезни. Необходимо также хранить всю соответствующую информацию о медицинском персонале. Система динамического архивирования позволит структурировать информацию о пациентах: фамилия, имя, отчество, адрес, дата рождения, перенесенные заболевания, дата поступления в больницу, противопоказания, результаты медицинских анализов, диагнозы. Аналогичным образом можно структурировать информацию о врачах и других сотрудниках. Система также позволит врачам обмениваться информацией о пациентах и записывать свои наблюдения.

Помимо этого система предоставляет возможность моментального сохранения документов, экономии места, удобный доступ к архивным документам благодаря реализации поиска по дескрипторам, возможность изменения сохраненных документов, обработку любой новой модели архивирования путем добавления классификаций и описаний из интерфейсов приложений без необходимости программных изменений.

Анализ существующих систем архивирования. Был изучен ряд приложений динамического архивирования (MFiles [4], BlueDoc [5], Speedy Organizer, IsoTracker [6] и др.) и выявлены их достоинства и недостатки (табл. 1).

Таблица 1

Сравнительный анализ систем управления документами

Элементы	MFiles	BlueDoc	Speedy Organizer	Paperport	IsoTracker
Роли / группы	–	–	+	–	+
Картотека / модуль	+	+	–	+	+
Почтовые группы	+	+	+	–	–
Пользователи	+	+	+	+	+
Роли менеджера, администратора, автора	–	–	+	–	+
Группы пользователей	+	+	+	+	+
Внутренние / внешние пользователи	+	+	–	–	+
Метаданные	–	–	–	–	+
Документы	–	–	–	–	–
Входящие / исходящие документы	–	–	–	–	–
Свойства	+	–	–	–	–
Внутренняя почта	–	–	–	–	–
Модули	–	–	–	–	+
Товар / группы товаров	–	–	–	–	–
Связи	–	–	–	–	–
Папки / подпапки	–	+	+	+	+

Главная необходимость заключается в том, что система должна позволять добавлять текст в графические изображения документов (скриншоты, фотографии) с возможностью их отображения как с добавленным текстом, так и без него, не дублируя изображение, чтобы сэкономить место, и не влияя на исходное изображение. Также система должна обеспечивать возможность оптического распознавания символов (OPC).

Постановка задачи. Необходимо разработать систему динамического архивирования, позволяющую:

1) создавать неограниченное множество классификаций в дополнение к любым шаблонам (дескрипторам) через интерфейсы и к специальной системе полномочий для пользователей;

2) осуществлять OPC для как можно большего количества языков;

3) добавлять электронную подпись к архивным документам для обеспечения их подлинности.

Основные компоненты системы показаны на рис. 1.

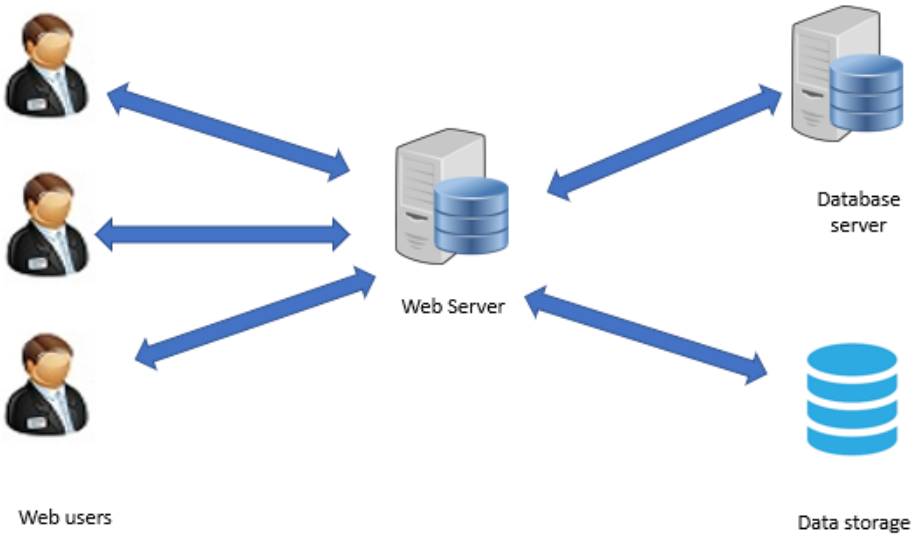

Рис. 1. Основные компоненты системы:

web-users — пользователи сети; web-server — веб-сервер; database server — сервер базы данных; data storage — хранение данных

Проектирование базы данных. База данных предназначена для создания неограниченного количества категорий в дополнение к неограниченному количеству атрибутов в соответствии с различными потребностями организаций. В качестве ядра базы данных использован программный комплекс Oracle [7].

Инфологическая модель показывает отношения наборов сущностей, хранящихся в базе данных. Сущность в данном контексте — это объект или компонент данных.

Инфологическая модель базы данных для системы динамического архивирования представлена на рис. 2.

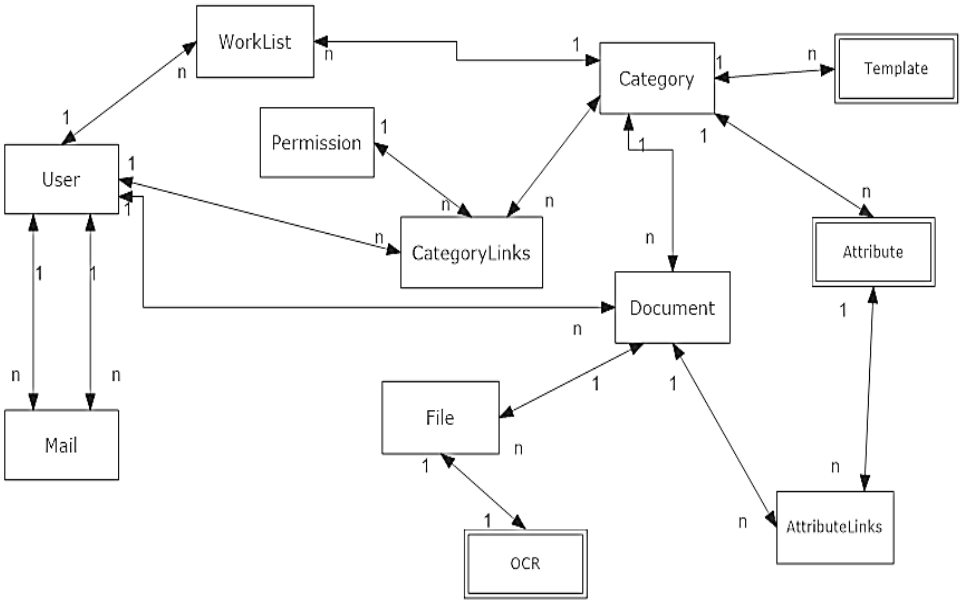

Рис. 2. Инфологическая модель базы данных:

worklist — рабочий список; user — пользователь; category — категория; mail — почта; categorylinks — ссылки по категориям; document — документ; attribute — атрибут; template — шаблон; permission — полномочия; file — файл; OCR — оптическое распознавание символов; attributelinks — ссылки на атрибуты

Даталогическая модель базы данных для системы динамического архивирования представлена на рис. 3.

База данных содержит информацию о следующих сущностях:

1) Attribute (атрибуты);
2) Categories (категории);
3) Document (документы);
4) DocumentAttribute (атрибуты документа);
5) DocumentFiles (файлы документа);
6) FileOCR (оптическое распознавание символов файла);
7) Permission (полномочия);
8) Template (шаблоны);
9) UserMail (почта пользователя);
10) UserPermissionCategory (категории пользователей по полномочиям);
11) Users (пользователи);
12) UserWorkList (рабочий список пользователей).

Разработка веб-приложения. Приложение разработано с использованием MVC (C#) [8] и состоит из двух основных частей:

1) специальный раздел для системного администратора, который создает необходимые категории с определением атрибутов категории, а также указанием полномочий пользователя (рис. 4);

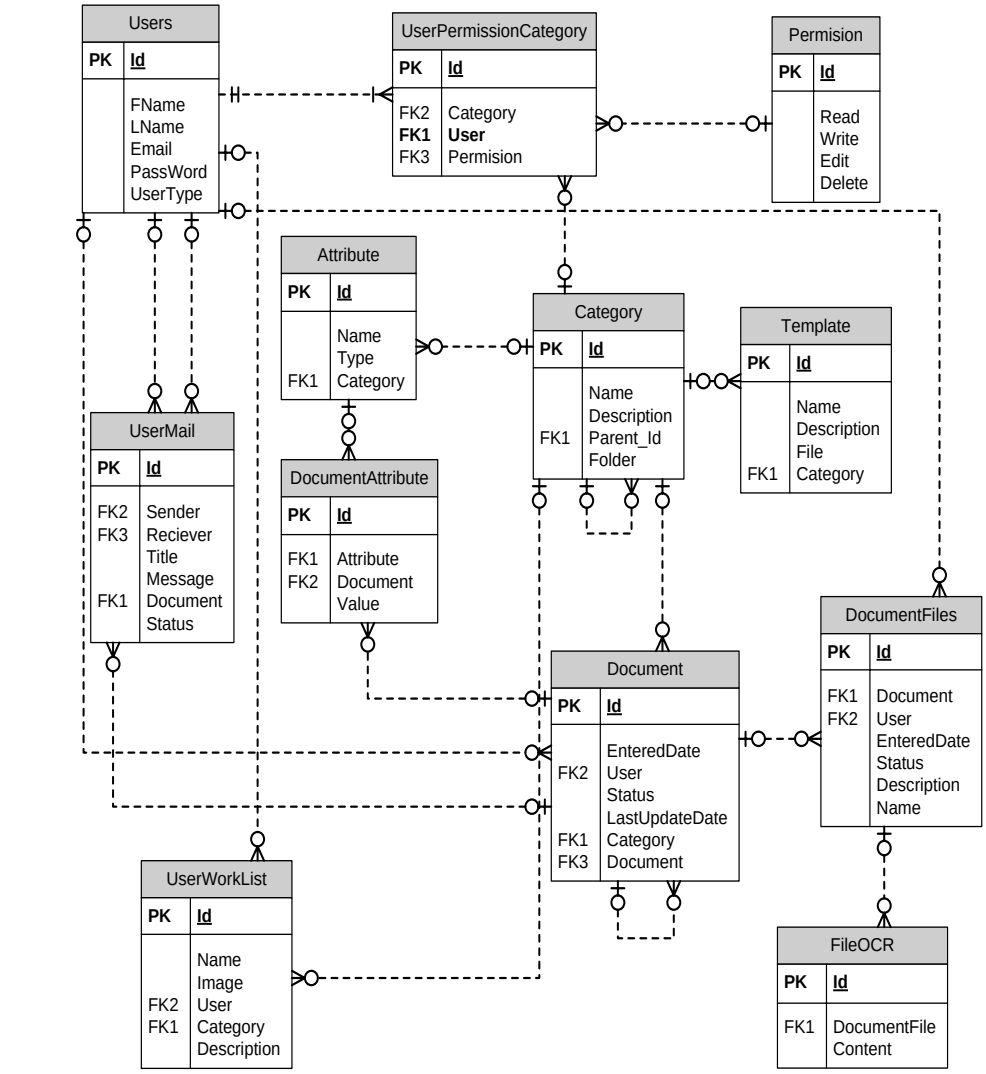

Рис. 3. Даталогическая модель базы данных

PK — персональный компьютер; id — идентификатор; FK1 — внешний; FK2 — внешний; FK3 — внешний; fname — имя; lname — фамилия; email — электронная почта; password — пароль; usertype — тип пользователя; read — читать; write — писать; edit — редактировать; delete — удалять; sender — отправитель; reciever — получатель; title — заголовок; message — сообщение; status — статус; name — название; type — тип; description — описание; parent_id — родительская страница; folder — папка; value — оценка, значение, стоимость; image — изображение; entereddate — дата ввода; lastupdate — дата последнего обновления; content — содержание

2) раздел пользователя для добавления новых документов (в том числе отсканированных версий, фотографий), атрибутов документов, а также получения текста по изображениям, добавления текста, поиска добавленных документы по их атрибутам, вывода на печать. Все перечисленное осуществляется в соответствии с заранее установленными системным администратором полномочиями для каждого пользователя (рис. 5).

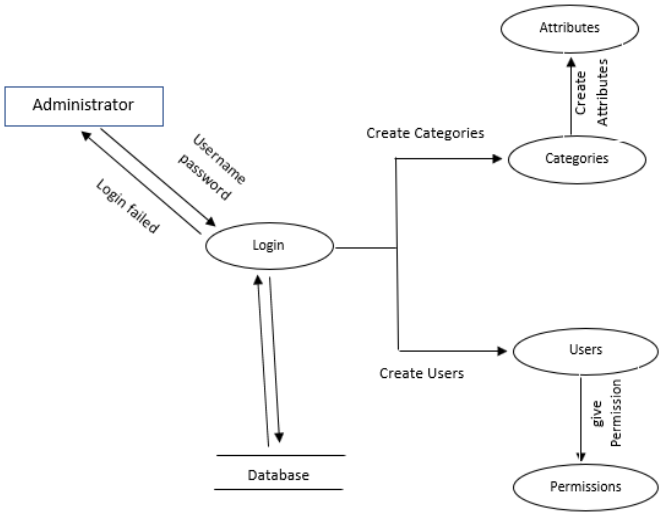

Рис. 4. Общая схема работы системного администратора:

administrator — администратор; login — вход, идентификация пользователя; database — база данных; username — имя пользователя; login failed — ошибка входа; create users — создание пользователей; create categories — создание категорий; create attributes — создание атрибутов; give permission — получение полномочий

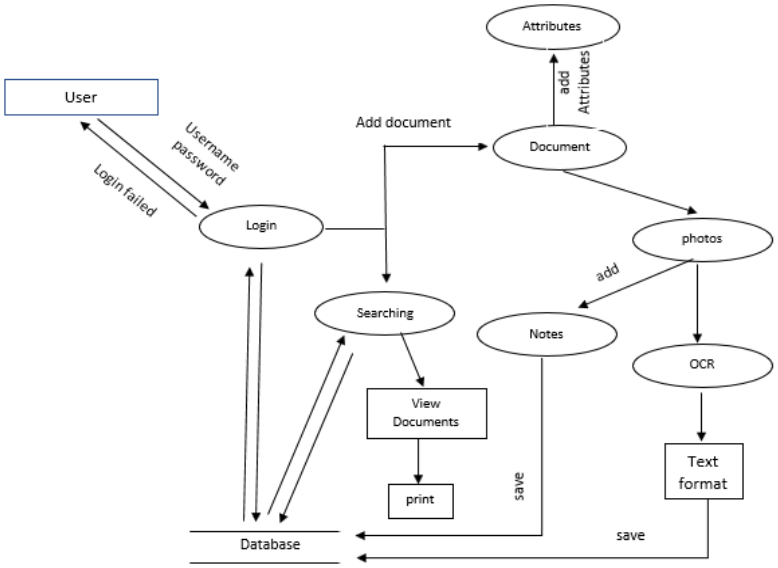

Рис. 5. Общая схема работы пользователя системы:

searching — поиск; view documents — просмотр документов; print — распечатать; notes — примечания; photos — фотографии; text format — текстовый формат; add document — добавление документа; add attributes —добавление атрибутов; add — добавление; save — сохранение

Оптическое распознавание символов. Оптическое распознавание символов — это использование технологии для различения печатных или рукописных символов текста по цифровым изображениям документов (например, отсканированному

бумажному документу). Процесс распознавания включает в себя изучение текста документа и перевод символов в код, который может быть использован для обработки данных. Процесс OPC иногда также называют распознаванием текста.

Результаты сравнительного анализа существующих систем OPC показаны в табл. 2.

Таблица 2

Сравнительный анализ систем OPC с открытым исходным кодом

Поддержка	Tesseract	CuneiForm	Microsoft A9T9
Онлайн-версия	–	–	–
Windows	+	+	+
MacOS	+	+	+
Linux	+	+	+
BSD	+	+	+
Языки программирования	C/C++	C/C++	C#
SDK	+	+	+
Количество языков	Более 100	28	21
Шрифты	Любые	Любые	Любые
Выходные форматы	txt, alto, hOCR, pdf	html, hOCR, rtf, txt	txt, doc, docx

В разрабатываемой системе динамического архивирования для OPC выбрана система Tesseract, поскольку она соответствует главному критерию: распознает символы более чем для 100 языков.

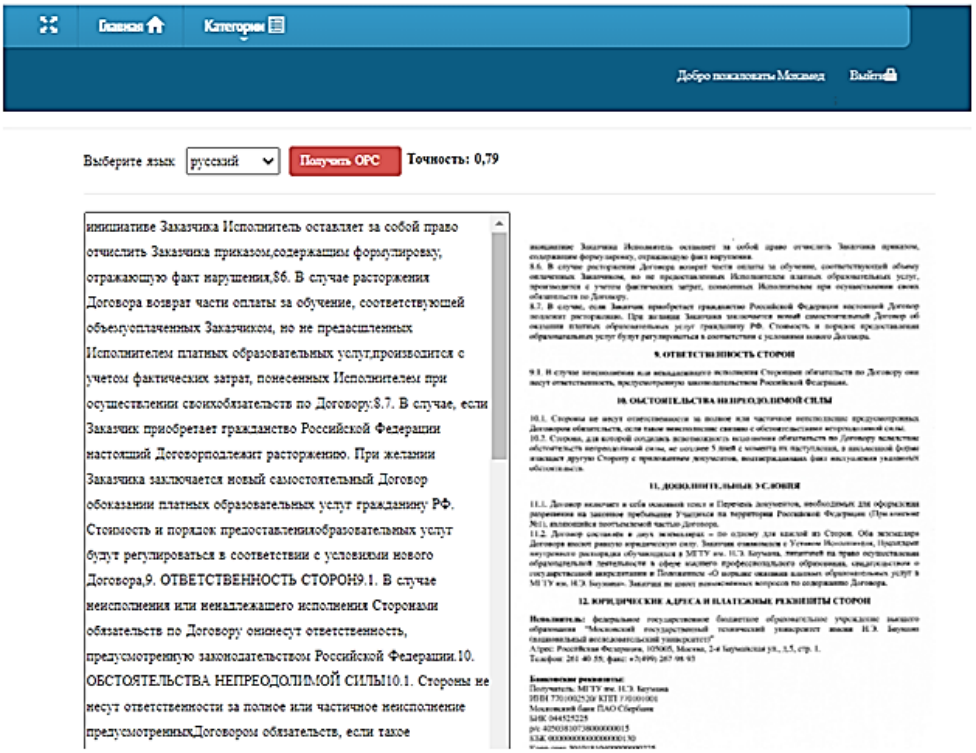

Рис. 6. Результат OPC документа

Tesseract — это самая известная программа OPC с открытым исходным кодом, разработанная компанией Hewlett-Packard. Это бесплатное программное обеспечение под лицензией Apache, которое спонсирует Google с 2006 г. Она считается одной из самых точных среди свободно доступных систем с открытым исходным кодом. В настоящее время Tesseract охватывает до 130 языков [9].

Пример OPC документа показан на рис. 6.

Добавление текста в изображение. При добавлении текста в изображение необходимо задавать множество параметров: тип шрифта, цвет текста, цвет и плотность контура, цвет и плотность тени и т. д. Для упрощения задачи была создана галерея текстовых шаблонов. Каждый текстовый шаблон включает в себя все перечисленные атрибуты, и при выборе пользователь может изменить только базовый цвет и размер шрифта.

Текст добавляется в изображение с помощью инструментов JavaScript и jQuery, (рис. 7).

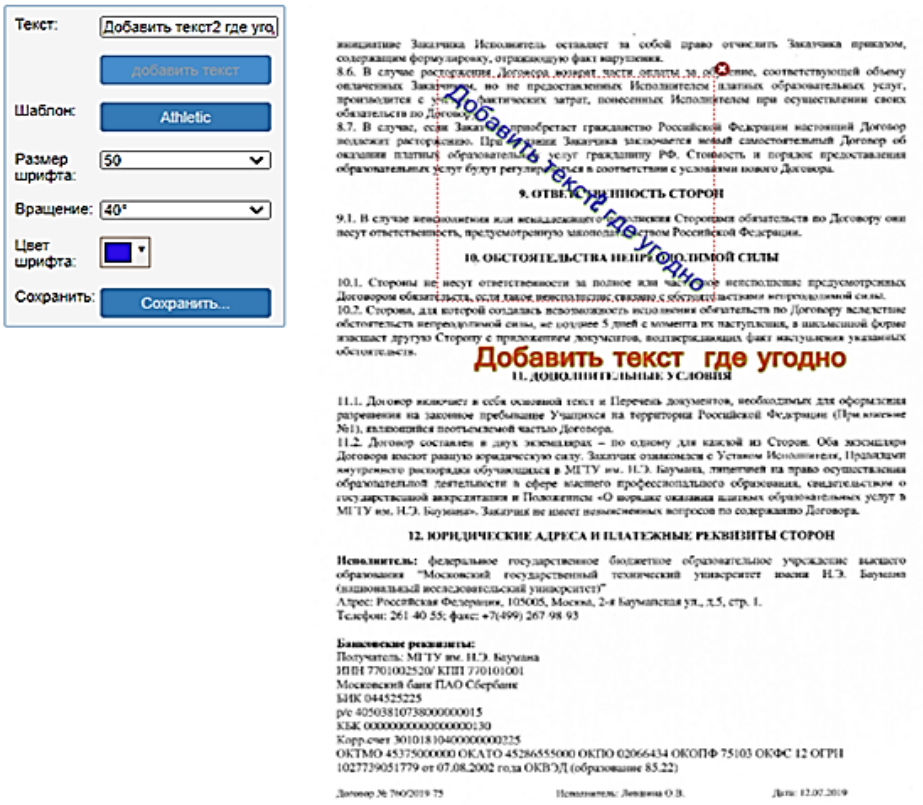

Рис. 7. Добавление текста в изображение

Выводы и перспективы дальнейшего развития. В статье представлен сравнительный анализ различных систем архивирования и выявлены характерные для таких систем недостатки и уязвимости. Разработана система динамического архивирования, обеспечивающая оптическое распознавание символов, электронную подпись документов и добавление текста в изображения.

Перспективы дальнейшего развития связаны с практическим внедрением разработанной системы в электронный документооборот, где необходима интеграция всего упомянутого выше функционала.

Литература

[1] Zantout H., Marir F. Document management systems from current capabilities towards intelligent information retrieval: an overview. *Int. J. Inf. Manage.*, 1999, vol. 19, no. 6, pp. 471–484. DOI: https://doi.org/10.1016/S0268-4012(99)00043-2

[2] Kirikova M. Flexibility of organizational structures for flexible business processes. *6th Workshop on Business Process*, 2005, pp. 123–130.

[3] Document management software. *capterra.com: веб-сайт*. URL: https://www.capterra.com/document-management-software (дата обращения: 15.10.2021).

[4] M-Files. *m-files.com: веб-сайт*. URL: https://www.m-files.com/en/latest-updates (дата обращения: 15.10.2021).

[5] BlueDoc document management system advantages. *blueproject.ro: веб-сайт*. https://www.blueproject.ro/bluedoc/advantages (дата обращения: 15.10.2021).

[6] Document management software. *isoTracker.com: веб-сайт*. URL: https://www.isotracker.com/products/document-management-software (дата обращения: 15.10.2021).

[7] Oracle database. *oracle.com: веб-сайт*. URL: https://www.oracle.com/database/ (дата обращения: 15.10.2021).

[8] ASP.NET MVC Pattern. *dotnet.microsoft.com: веб-сайт*. URL: https://dotnet.microsoft.com/apps/aspnet/mvc (дата обращения: 15.10.2021).

[9] Tesseract-OCR. *github.com: веб-сайт*. URL: https://github.com/tesseract-ocr/tesseract/ (дата обращения: 15.10.2021).

Алия Мухаммад — студентка магистратуры кафедры «Компьютерные системы и сети», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация.

Научный руководитель — Миков Дмитрий Александрович, кандидат технических наук, доцент кафедры «Компьютерные системы и сети», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация.

Ссылку на эту статью просим оформлять следующим образом:
Алия Мухаммад. Разработка системы динамического архивирования. *Политехнический молодежный журнал*, 2021, № 12(65). http://dx.doi.org/10.18698/2541-8009-2021-12-754

DEVELOPMENT OF A DYNAMIC ARCHIVING SYSTEM

Aliah Muhammad eng.muhammadaliah@gmail.com
SPIN-code: 4669-7154

Bauman Moscow State Technical University, Moscow, Russian Federation

Abstract

Document management is an important part of the work of organizations of all levels in various fields of activity. It includes the management of content, digital assets, photocopying and document flow. The article describes the process of developing a dynamic archiving system that allows one to create categories of documents and their attributes, carry out optical character recognition and, as a result, increase the efficiency of document management. With its help, companies will be able to build an optimal document management model, regardless of their attributes, content, time and storage location. The goal is achieved thanks to the ability to search documents by the tags specified for them or by their content, providing optical character recognition. In addition, a model for differentiating user access has been developed, and the ability to add text to image files has been provided.

Keywords

Dynamic archiving system, optical character recognition, footnotes, documents, photo editing, dialog model, database, role-based access control, computer systems

Received 13.05.2021

References

[1] Zantout H., Marir F. Document management systems from current sapabilities towards intelligent information retrieval: an overview. *Int. J. Inf. Manage.*, 1999, vol. 19, no. 6, pp 471–484. DOI: https://doi.org/10.1016/S0268-4012(99)00043-2

[2] Kirikova M. Flexibility of organizational structures for flexible business processes. *6th Workshop on Business Process*, 2005, pp 123–130.

[3] Document management software. *capterra.com: website*. URL: https://www.capterra.com/document-management-software (accessed: 15.10.2021).

[4] M-Files. *m-files.com: website*. URL: https://www.m-files.com/en/latest-updates (accessed: 15.10.2021).

[5] BlueDoc document management system advantages. *blueproject.ro: website*. https://www.blueproject.ro/bluedoc/advantages (accessed: 15.10.2021).

[6] Document management software. *isoTracker.com: website*. URL: https://www.isotracker.com/products/document-management-software (accessed: 15.10.2021).

[7] Oracle database. *oracle.com: website*. URL: https://www.oracle.com/database/ (accessed: 15.10.2021).

[8] ASP.NET MVC Pattern. *dotnet.microsoft.com: website*. URL: https://dotnet.microsoft.com/apps/aspnet/mvc (accessed: 15.10.2021).

[9] Tesseract-OCR. *github.com: website*. URL: https://github.com/tesseract-ocr/tesseract/ (accessed: 15.10.2021).

Muhammad Aliah — Student, Department of Computer Systems and Networks, Bauman Moscow State Technical University, Moscow, Russian Federation.

Scientific advisor — Mikov D.A., Cand. Sc. (Eng.), Assoc. Professor, Department of Computer Systems and Networks, Bauman Moscow State Technical University, Moscow, Russian Federation.

Please cite this article in English as:
Muhammad Aliah. Development of a dynamic archiving system. *Politekhnicheskiy molodezhnyy zhurnal* [Politechnical student journal], 2021, no. 12(65). http://dx.doi.org/10.18698/2541-8009-2021-12-754.html (in Russ.).