УДК 004.42

РАЗРАБОТКА ИНТЕРНЕТ-КАТАЛОГА ДЛЯ ОРГАНИЗАЦИИ ДОСТУПА К КОРПОРАТИВНОМУ ХРАНИЛИЩУ ЭЛЕКТРОННЫХ ДОКУМЕНТОВ ТПУ

Шерстнёв В.С., Распопов А.В.

ФГБОУ ВПО «Национальный исследовательский Томский политехнический университет», Томск, Россия (634050, г. Томск, пр. Ленина, 30), e-mail: vss@tpu.ru

В данной статье описываются этапы создания интернет-ресурса для публикации электронных документов из корпоративного хранилища университета, построенного с использованием программного продукта Oracle Universal Content Management. Одной из целей создания подобного интернет-ресурса является необходимость сделать документы корпоративного электронного хранилища доступными для их обнаруживаемости и полнотекстовой индексации современными поисковыми системами (Google.com, Yandex.ru). В статье описывается существующая архитектура созданного варианта хранилища электронных документов университета, рассматриваются возможные архитектурные варианты создания интернет-каталога, указаны возможные протоколы обмена данными между хранилищем документов и приложением интернет-каталога, приведены этапы и особенности проектирования системы, а также приведены результаты выполненной реализации. В работе описаны итоги пробной эксплуатации разработанного интернет-каталога, описаны перспективы его дальнейшего развития.

Ключевые слова: интернет-доступ, электронные документы, корпоративное хранилище, Oracle UCM, ASP.NET, MVC, полнотекстовый поиск, поисковые системы, публикация контента, разработка приложения, взаимодействие с Oracle UCM, ТПУ.

DEVELOPMENT OF THE INTERNET CATALOGUE FOR THE ORGANIZATION OF ACCESS TO CORPORATE DATA WAREHOUSE OF ELECTRONIC DOCUMENTS OF TPU

Sherstnev V.S., Raspopov V.S.

National Research Tomsk Polytechnic University, Tomsk, Russia (634050, г. Томск, Lenin Avenue, 30), e-mail: vss@tpu.ru

This article describes the steps for creating the online resource for publishing electronic documents from a corporate repository of the University, which was built with the use of a software product Oracle Universal Content Management. One of the goals of this online resource is the need to make corporate electronic document repository accessible to their detectability and full-text indexing of modern search engines (Google.com, Yandex.ru). This paper describes the architecture of the existing version of the created repository of electronic documents university, possible architectural options for creating an online resource, the possible communication protocols between repository and the application documents online resource are given stages and features of the system design and the result of the implementation. This paper describes the results of the test run of the developed online resource describes the prospects for its further development.

Keywords: Internet access, electronic documents, corporate repository, Oracle UCM, ASP.NET, MVC, full-text search, the search engines, the publication of content, application development, interaction with Oracle UCM, TPU.

Введение. Безусловно, что создание и поддержка в актуальном состоянии корпоративного хранилища электронных документов является важной задачей [3]. Тем не менее после создания соответствующих структур для хранения электронных документов появляется задача об удобном доступе к ним и обнаруживаемости их поисковыми системами.

В Томском политехническом университете (ТПУ) в качестве пилотного проекта разрабатывается система электронного хранилища документов на основе программного продукта Oracle Universal Content Management (Oracle UCM) [1]. Сам продукт Oracle UCM в своей работе использует СУБД Oracle для хранения электронных документов и их

полнотекстовой индексации, а также набор стандартных пользовательских веб-интерфейсов и веб-сервисов для взаимодействия с хранилищем электронных документов, как представлено на рис. 1.

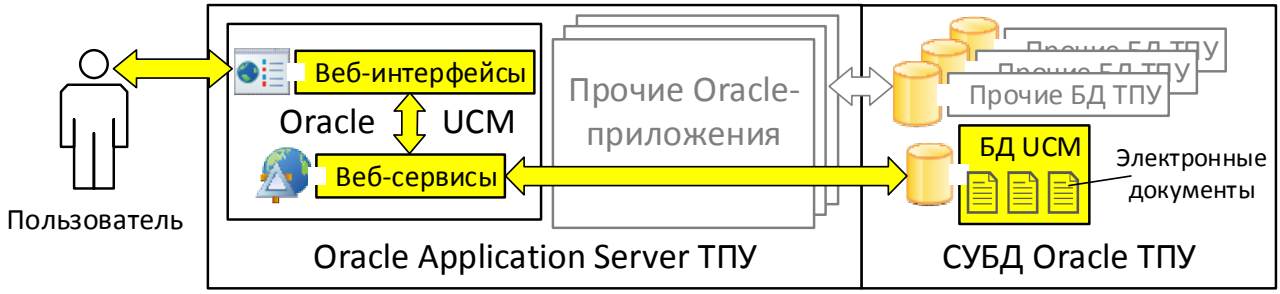

Рис. 1. Обобщённая архитектура Oracle UCM в инфраструктуре ТПУ

В традиционном варианте доступ пользователей к электронным документам Oracle UCM производится по траектории, выделенной на схеме желтым цветом. При этом задействованы стандартные пользовательские веб-интерфейсы и сервисы Oracle UCM.

Цель исследования

Варианты решения. С одной стороны, для построения интернет-каталога поверх хранилища электронных документов в корпоративной сети ТПУ возможно было бы воспользоваться существующими возможностями Oracle UCM, как показано на рис. 1. При этом дизайн пользовательского интерфейса интернет-каталога и его функционал ограничивались бы возможностями компонент UCM.

Данный вариант был рассмотрен и отвергнут, как непригодный с точки зрения дальнейшего расширения функционала. Готовые пользовательские интерфейсы Oracle UCM являются стандартными по дизайну и функциям, что не всегда удобно для встраивания в существующие корпоративные системы.

С другой стороны, для построения интернет-каталога электронных документов возможно разработать свой пользовательский веб-интерфейс, взаимодействующий напрямую с базовыми веб-сервисами Oracle UCM. Пример интернет-каталога электронных документов, построенного по второму варианту архитектуры, показан на рис. 2.

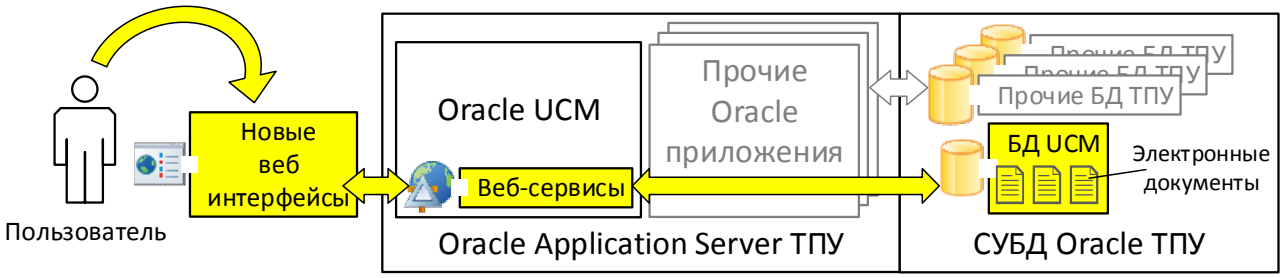

Рис. 2. Вариант построения интернет-каталога с использованием нового пользовательского веб-интерфейса

Такой вариант позволит легче изменять и расширять функционал интернет-каталога в дальнейшем, по новому осуществляя обработку информации, извлечённой из хранилища электронных документов. Отличительной особенностью этого варианта построения интернет-каталога является возможность беспрепятственной модернизации его бизнес-логики и дизайна. Такой вариант авторам работы видится более предпочтительным, так как предоставляет большую свободу действий в дальнейшем наращивании возможностей интернет-каталога.

Проектирование. В процессе проектирования выбранного варианта были определены детали внутренней архитектуры интернет-каталога и принципы взаимосвязи с корпоративным хранилищем электронных документов. Для взаимодействия с хранилищем документов возможны несколько протоколов взаимодействия: Remote IDC (RIDC) [5], Simple Object Access Protocol (SOAP) [8]. Протокол RIDC является проприетарной разработкой Oracle и, разумеется, обладает большими возможностями по взаимодействию с Oracle UCM. Но Oracle предоставляет программные библиотеки с высокоуровневыми функциями RDIC только для языка Java. Второй доступный для взаимодействия протокол (SOAP) – является отлично документированным международным стандартом [8] и обладает открытыми готовыми библиотеками на многих языках программирования высокого уровня (PHP, C#, Java, Pyton и т.д.). Вследствие этого в работе был использован именно протокол SOAP. Использование SOAP определило сервис-ориентированный (SOA, Service-Oriented Architecture [7]) тип архитектуры интернет-каталога. Архитектура SOA характерна слабым связыванием компонент, повторным использованием сервисов и высокой расширяемостью. Схема общего взаимодействия проектируемого веб-приложения интернет-каталога, сервера Oracle UCM и глобальных поисковых систем представлена на рис. 3.

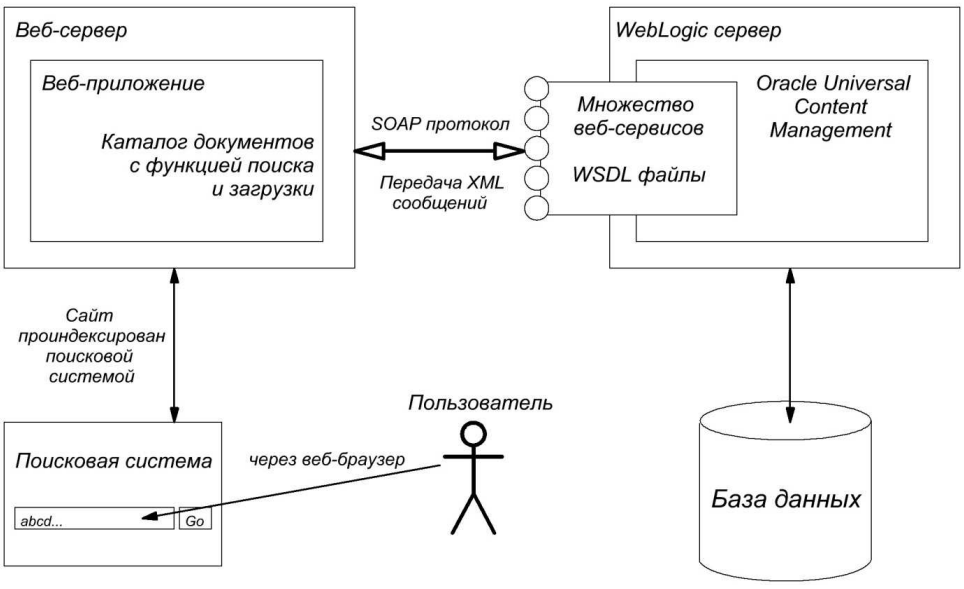

Рис. 3. Структура взаимоотношений компонент системы

На основе проведенного анализа веб-сервисов Oracle UCM установлено, что для работы интернет-каталога достаточным будет использование трех базовых веб-сервисов Oracle UCM: DOC_INFO, GET_FILE, GET_SEARCH_RESULTS, позволяющих получать информацию о документе, загружать документ на компьютер клиента и производить поиск документа по требуемым атрибутам, соответственно.

В процессе проектирования интернет-каталога использован унифицированный язык моделирования UML (спецификация UML 2.4.1, принятая ISO/IEC 19501:2005), с помощью которого построены диаграмма прецедентов (вариантов использования) и диаграмма классов [9].

Реализация. Веб-приложение интернет-каталога реализовано при помощи ASP.NET MVC фреймворка на языке C# с использованием Visual Studio 2012.

На рис. 4 представлен фрагмент страницы с документами, найденными при выполнении полнотекстового поискового запроса с ключевым словом «шаблон». Значимым является то, что искомое ключевое слово было найдено именно в содержимом документов, а не в их метаописании (заголовке, списке авторов, кратком описании и т.п.).

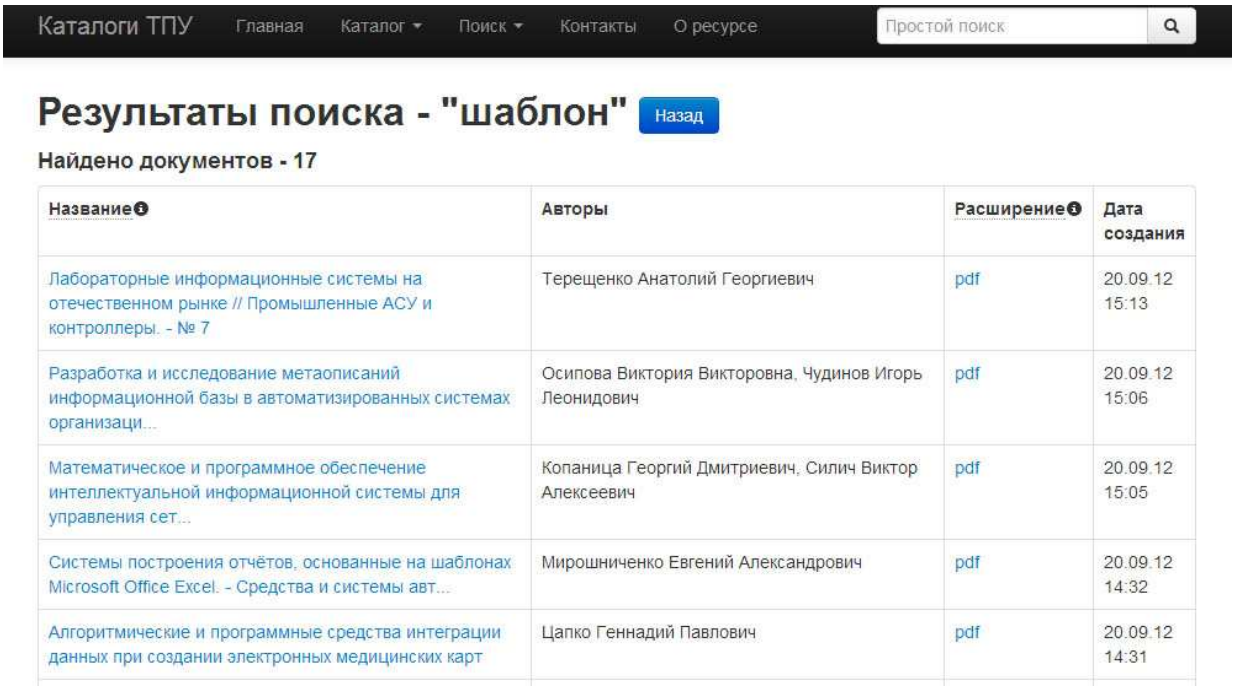

Рис. 4. Фрагмент результатов полнотекстового поискового запроса по ключевому слову «шаблон»

При реализации интернет-каталога использовалась концепция MVC (Model-View-Controller), согласно которой приложение было разделено на три области: модель данных, представления и контроллеры. В рамках среды разработки Visual Studio 2012 решение состоит из двух частей-проектов. Один из проектов представляет собой библиотеку классов C# и отвечает за предоставление доступа к модели данных. Второй проект отвечает за пользовательский интерфейс клиентской части и содержит представления для визуализации

информации, пользовательского интерфейса, а также контроллеры для связи между моделью и представлением. В целом в процессе кодирования интернет-каталога были реализованы: модель данных, 12 представлений (с использованием рабочего набора библиотеки Razor [6] и дополнительного пакета twitter.bootstrap), 3 контроллера.

Пробная эксплуатация. Разработанный интернет-каталог был развернут для тестирования и пробной эксплуатации на выделенном сервере Томского политехнического университета [2] под управлением операционной системы Windows 7 и веб-сервера Internet Information Server v.7.0. Для проверки обнаруживаемости контента интернет-каталога из внешних поисковых систем ресурс был поставлен в очередь на индексирование в поисковых системах Google и Яндекс с помощью соответствующего программного инструментария [4].

В течение одной недели поисковая система Google внесла в свою индексную базу данных более 500 HTML-страниц из установленного интернет-каталога. Каждая из проиндексированных страниц являлась запросом на поиск документов по автору или названию. В связи с этим информация о документах интернет-каталога обнаруживается посредством сервиса Google. На рис. 5 приведены результаты поискового запроса по ключевой фразе «ГИС модуль для информационной системы агрохимического», где видно, что ресурс catalog1.vt.tpu.ru выдается поисковым механизмом Google на первом месте среди примерно 5540 результатов.

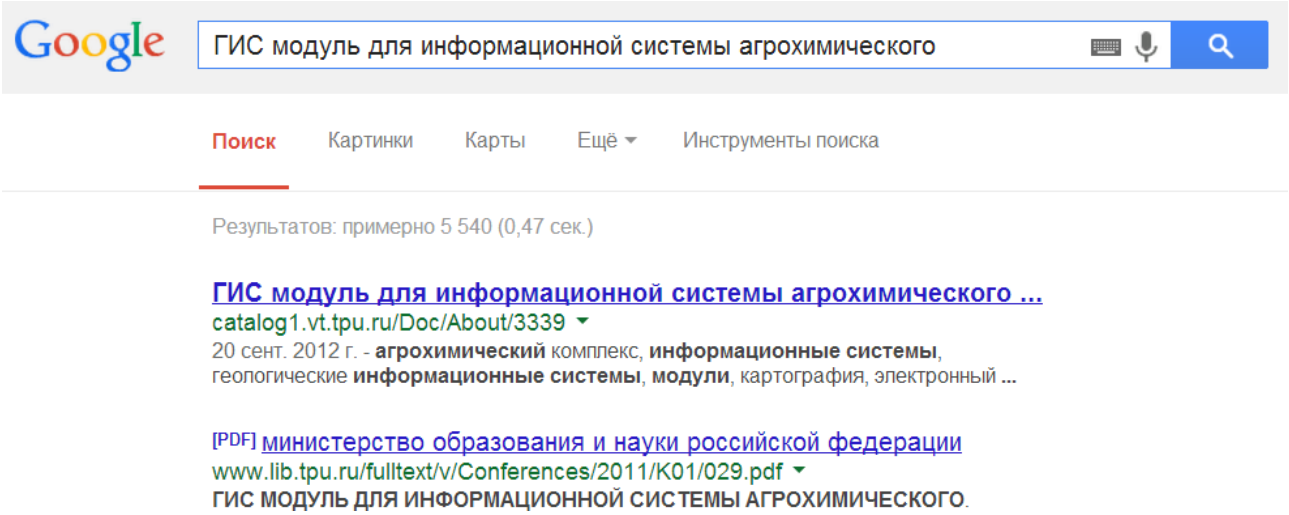

Рис. 5. Результаты поисковой системы Google на запрос «ГИС модуль для информационной системы агрохимического»

Разработанный программный код интернет-каталога электронных документов не является полностью законченным. Перспективой модернизации ресурса является повышение его быстродействия и отказоустойчивости, перехват инъекций программного кода в поисковых запросах, дальнейшее продвижение интернет-ресурса в рейтингах поисковых систем.

Список литературы

1. Oracle UCM // КОРУС Консалтинг [Электронный ресурс]. – Режим доступа: http://ecm.korusconsulting.ru/technology/oracleucm/ (дата обращения: 01.11.2013).
2. Томский политехнический университет. Каталоги ТПУ [Электронный ресурс]. – Режим доступа: http://catalog1.vt.tpu.ru (дата обращения: 01.11.2013).
3. Шерстнёв В.С., Иванов С.С., Акулин И.А. Использование Oracle Universal Content Managment в качестве корпоративного хранилища документов ТПУ // Вестник науки Сибири : электронный научный журнал / Томский политехнический университет (ТПУ). — 2011. — № 1 (1). — С. 302-307.
4. Google developers // Google for Webmasters - Webmaster EDU [Электронный ресурс]. – (дата обновления: 11.04.2012) – Режим доступа: https://developers.google.com/webmasters/googleforwebmasters/?hl=de (дата обращения: 01.11.2013).
5. Remote Intradoc Client (RIDC) Developer Guide // Oracle [Электронный ресурс]. – Режим доступа: http://docs.oracle.com/cd/E10316_01/ContentIntegration/ridc/ridc-developer-guide.pdf, свободный (дата обращения 01.11.2013).
6. ASP.NET Razor view engine // Wikipedia [Электронный ресурс]. – Режим доступа: http://en.wikipedia.org/wiki/ASP.NET_Razor_view_engine (дата обращения 01.11.2013).
7. Reference Model for Service Oriented Architecture 1.0 // OASIS [Электронный ресурс]. – Режим доступа: https://www.oasis-open.org/committees/download.php/19679/soa-rm-cs.pdf (дата обращения: 01.11.2013).
8. SOAP Version 1.2 Part 1: Messaging Framework (Second Edition) // World Wide Web Consortium [Электронный ресурс]. – Режим доступа: http://www.w3.org/TR/soap/ (дата обращения: 01.11.2013).
9. UML 2.4. Documents Associated With Unified Modeling Language (UML), V2.4 // OMG [Электронный ресурс]. – Режим доступа: http://www.omg.org/spec/UML/2.4/ (дата обращения: 01.11.2013).

Рецензенты:

Ким В.Л., д.т.н., профессор кафедры вычислительной техники Института кибернетики ФГБОУ ВПО «НИ ТПУ», г. Томск.

Авдеева Д.К., д.т.н., профессор, директор ООО «Медприбор», г. Томск.