УДК 004.413.2

ТЕХНОЛОГИЧЕСКИЙ ЦИКЛ РАЗРАБОТКИ ПОИСКОВОЙ СИСТЕМЫ, АГРЕГИРУЮЩЕЙ ЦИТАТЫ ИЗ КНИГ

Р. В. Мосоло́в[0000-0002-4399-4397]

Корпорация X5 Retail Group (Москва)

R.V.Mosolov@ya.ru

Аннотация

Описан технологический цикл разработки поисковой системы по 14 книгам философской направленности Л.А. Секлитовой и Л.Л. Стрельниковой, состоящий из 6 этапов работ. Идеи статьи могут быть полезны при проектировании и разработке программного обеспечения, агрегирующего цитаты из серий книг, монографий, научных публикаций или научно-периодических изданий, например, в целях хранения персональных ссылок на вторичные источники, часто пригождающиеся при написании научных статей и оформлении презентаций в педагогике. Поисковая система является результатом года работ автора и группы из примерно 30 волонтёров. Система представляет собой сервис, встроенный в веб-приложение. Технологический стек: Jade, CSS, JS, Node.js, Express.js, ESLint, Jest.

Ключевые слова*: цитаты писателей, агрегатор книг, цитаты философов, поисковые системы книги, цитаты из литературы, философские цитаты, оптимизация и продвижение в поисковых системах книг, лучшие цитаты из книг, разработка поисковой системы.*

ВВЕДЕНИЕ

В мае 2021 года в рамках деятельности региональной общественной организации «Центр Духовного Развития Человека «Золотая Раса»» (далее – ЦЗР; https://www.rusprofile.ru/id/1217700649473) была начата разработка сервиса, агрегирующего цитаты из 79 книг Л.А. Секлитовой и Л.Л. Стрельниковой (https://gold-race.org/searcher). Автор статьи отвечал за проектирование и программирование алгоритмов данного сервиса, разрабатывал как клиентскую (см. рис. 1), так и серверную части сервиса.

Цель создания сервиса – помочь пользователям найти ответ на философский вопрос при вводе определённого запроса в поисковую строку данного сервиса. Первый коммит сервиса был сделан 15 мая 2021 года (https://github.com/R-Mosolov/spircent/tree/7c190f7bc9919a50816001a230a87b2b2b23836d). В целях функциональной безопасности сервиса, а также защиты авторских прав авторов цитат исходный код сервиса хранится в приватном репозитории.

Все работы волонтёрами велись на благотворительной основе и, как правило, в свободное от работы время. В среднем один волонтёр был готов посвящать работам в Центре около 8 часов в неделю.

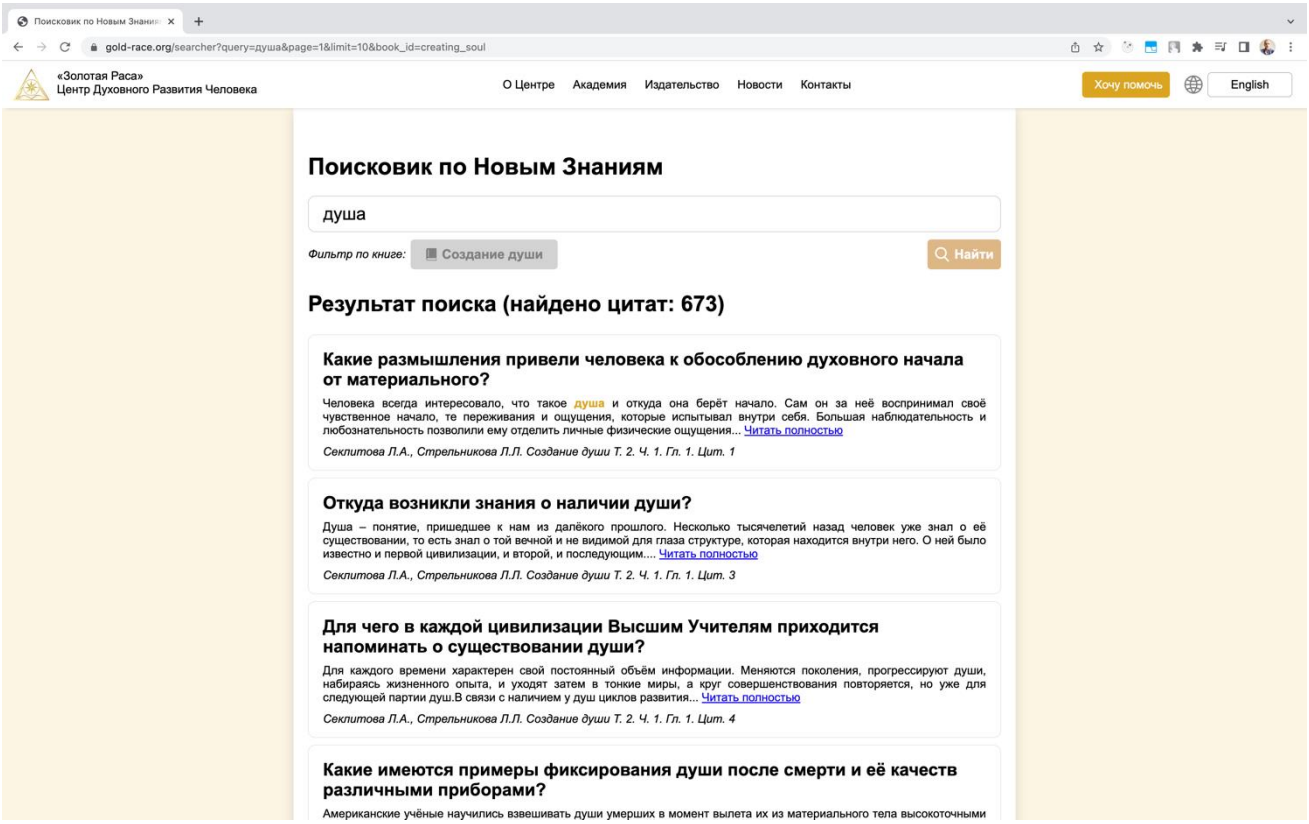


Рисунок 1. Пользовательский интерфейс сервиса по поиску книжных цитат

ТЕХНОЛОГИЧЕСКИЙ СТЕК

Следует отметить, что сами разработанные поисковые алгоритмы, с нашей точки зрения, не представляют существенной сложности, поскольку основная часть усилий была сосредоточена, скорее, на создании руководств пользователей, организационных условий для переноса цитат на сайт, а также обеспечении жизнеспособности IT-инфраструктуры.

В таблице ниже представлен список основных технологий, использованных при разработке сервиса. Отметим, что это неполный список того, чем автору статьи довелось заниматься в процессе разработки сервиса. Также в силу благотворительного характера работ в организации, способствовавшей привлечению и удержанию программистов узкого профиля (речь идёт о разработчиках серверной части и DevOps), автору требовалось освоить работу с циклом CI/CD. Если в первый месяцы существования веб-приложения данную настройку можно было спокойно делегировать PaaS-платформе Heroku (https://heroku.com), то после введения ряда санкций в отношении России, связанных с военными действиями на Украине (https://habr.com/ru/post/653605/) и как следствие приведших к появлению «Указа о мерах по обеспечению ускоренного развития отрасли информационных технологий в России», введённого Президентом В.В. Путиным (http://kremlin.ru/acts/news/67893), потребовалось осваивать, как вручную поднять и настроить виртуальный приватный сервер (VPS), чтобы перенести веб-приложение на российский сервер для обеспечения дальнейшей работоспособности созданного сервиса. Потребность эта особенно остро проявилась тогда, когда из-за ошибок, появляющихся в интерфейсе Heroku, стало невозможным публиковать какие-либо изменения на сайте. Предполагаем, что последнее было связано с поломкой интеграции Heroku с GitHub. Хотя руководство GitHub публично утверждало, что будет защищать права отдельно взятых разработчиков на свободное распространение кода независимо от страны их проживания (https://github.blog/2022-03-02-our-response-to-the-war-in-ukraine/), факт поломки, коррелирующий с политическими событиями, свидетельствовал об обратном. В итоге веб-приложение было перенесено на сервер под ОС Ubuntu 20.04 LTS (3 Гб ОЗУ, 3 ядра, 60 Гб SSD).

Таблица 1. Технологический стек сервиса «Поисковик по Новым Знаниям»

№	**Название технологии**	**Тип технологии**
1	Jade	Препроцессор
2	CSS	Язык стилизации
3	JS	Язык программирования

4	Node.js	Серверное окружение
5	Express.js	Серверный фреймворк

ТЕХНОЛОГИЧЕСКИЙ ЦИКЛ ПЕРЕНОСА ЦИТАТ

Перенос цитат состоял из 5–6 этапов работ (см. рис. 2). Вариативность с одним этапом была обусловлена тем, что работающий над переносом цитат волонтёр мог либо использовать Эксель-таблицы, либо отказаться от них. Эксель был выбран не просто так. Благодаря жёсткой структуре ячеек, Эксель-таблицы быстро преобразуются в формат JSON – основной формат хранения цитат в проекте. Однако, к нашему сожалению, выяснилось, что для большинства волонтёров работа с Экселем представляла техническую сложность. Поскольку проект реализовывался на благотворительной основе, то для его осуществления привлекались работники разных профессий, имеющие сильно различающийся уровень владения компьютером, и разных возрастов.

Приведём некоторые цифры:

– формально в проекте числилось около 20–30 волонтёров на разных стадиях его реализации;

– по факту, регулярно работало около 6–7 волонтёров;

– цитаты в Эксель-таблице прислал только 1 человек.

Предпочтение Ворда Экселю отдавалось не только из-за технических сложностей при переносе цитат. Также оно было связано с тем, что Эксель (по крайней мере, по умолчанию) не проверяет орфографию и пунктуацию. Последние являются критическим бизнес-требованием при работе с книгами, поскольку вопросы для цитат составлялись самими волонтёрами, а не брались в готовом виде из книг.

Ближе к завершению работ над сервисом, когда параллельно велись работы по интернационализации интерфейса сайта, нами эмпирически была обнаружена *«золотая середина» между Экселем и Вордом*, позволяющая, с одной стороны, использовать жёсткую систему ячеек таблиц, с другой стороны, включающая проверку орфографии и пунктуации в текстах. Данной «золотой серединой» стало заполнение текстов прямо в Ворде, но с предварительным созданием таблицы в нём.

Остановимся на подборе ключевых слов. Поскольку архитектурно сайт, на котором расположен сервис, задумывался с целью поисковой оптимизации его страниц, или SEO (отсутствие большого маркетингового бюджета – критический фактор для региональной общественной организации), что влияло на выбор его технологического стека, то для его страниц требовалось подбирать ключевые слова. Изначально мы планировали поручить данную задачу самим волонтёрам, чтобы снизить нагрузку на программистов, но, столкнувшись с тем, что для большинства из них освоение Экселя представляло сложность, мы решили отказаться от этой идеи и сосредоточить данную компетенцию в рамках Отдела Информационных Технологий организации. К сожалению, и это оказалось неверным решением, поскольку впоследствии, в процессе наращивания компетенции по SEO, мы узнали, что «поисковые пауки» не заполняют форм (тогда как её заполнение необходимо для отправки поискового запроса браузеру) [1, с. 245].

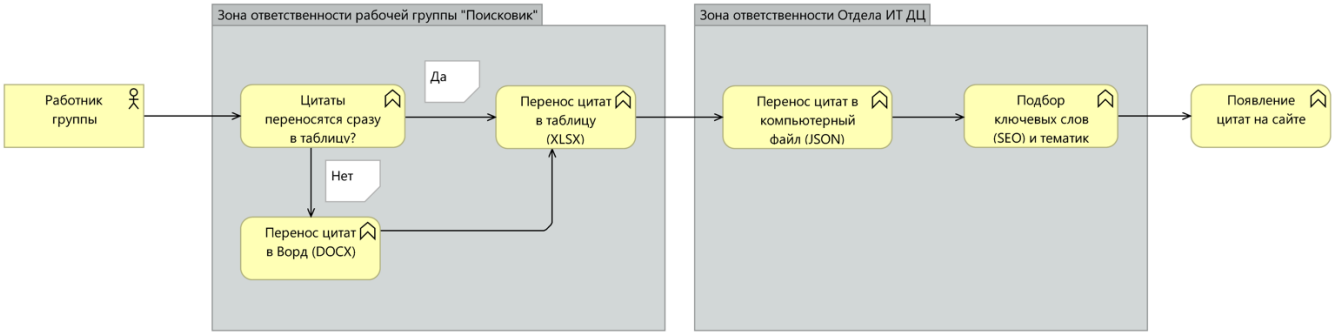


Рисунок 2. Блок-схема технологического цикла переноса цитат

ОРГАНИЗАЦИОННЫЕ И ТЕХНОЛОГИЧЕСКИЕ ОШИБКИ

Отметим ряд ошибок, совершённых на технологическом и организационном уровнях работы над проектом. Возможно, эта информация поможет разработчикам/заказчикам аналогичных сервисов избежать их впоследствии, учась на опыте других.

Первая ошибка (организационная) происходила из мотивационной основы, в рамках которой работали составители цитат. Дело в том, что книги делились на цитаты исключительно на благотворительной основе, что создавало ряд ограничений в механизмах мотивации работников. Так, например, у нас не было столь распространённых в России рычагов отсылки к окладу или премии. В связи

с этим была совершена одна из наиболее крупных, на наш взгляд, ошибок. Постановка задачи волонтёрам в первые 1–3 месяца работ утрированно звучала следующим образом: «Вот вам книга, у Вас есть год, будем ожидать вашего возвращения, когда будут готовы все цитаты».

Подобная постановка задачи впоследствии могла привести к осуществлению на практике «каскадной модели» управления проектом [2, с. 46]. Хотя такая модель использовалась бы не для управления группой разработчиков, а для управления специалистами в предметной области, тем не менее, как показала практика переноса цитат в первый месяц работ, она была не столь эффективна, как если бы мы использовали отдельные аспекты гибких (Agile) методологий, например, SCRUM.

В связи с этим в дальнейшем было принято **решение** разделить каждую книгу на смысловые части и просить волонтёров отправлять промежуточные результаты работ, тем самым распределяя объёмы работ по смысловым частям книги (главам), или спринтам (см. рис. 3). Это позволило сократить срок между началом работ волонтёров и выводом сервиса в продакшен примерно в 4 раза.

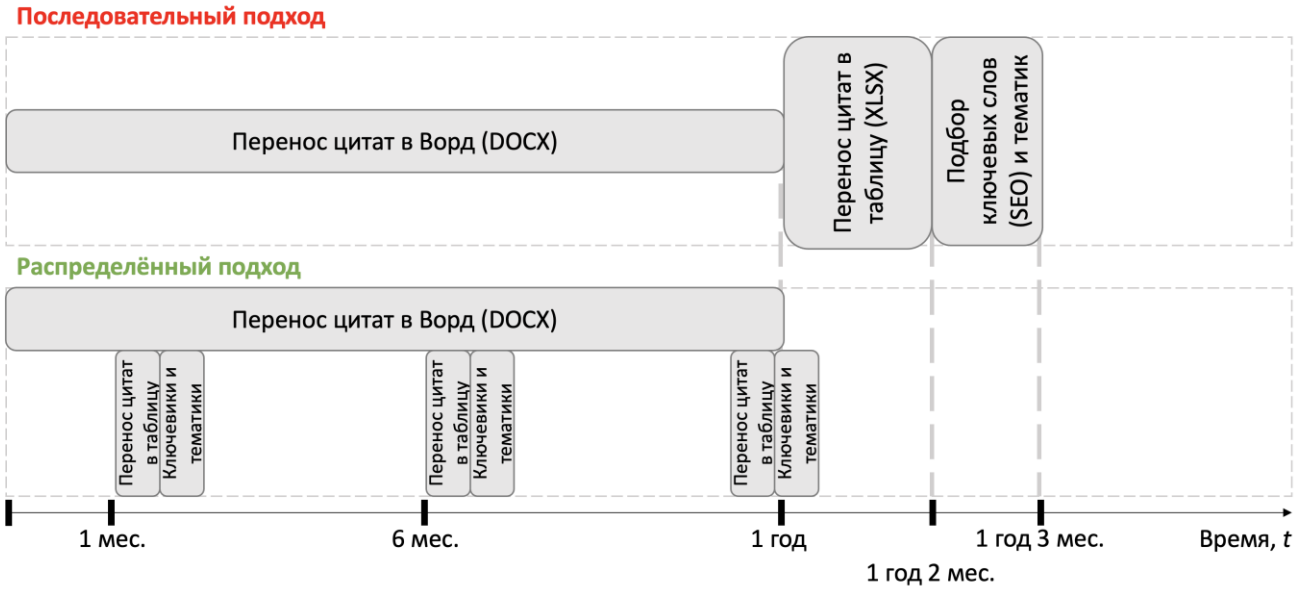


Рисунок 3. Различия последовательного и распределённого подходов при разделении книг на цитаты

Вторая ошибка (технологическая) заключалась в том, что поскольку сервис разрабатывался для региональной общественной организации, то последнее

накладывало существенный отпечаток на финансовую составляющую. Так, например, организация существует на членские взносы и пожертвования, что приводит к относительно нестабильному рекламному бюджету, а также, в принципе, малости его доли, могущей быть выделенной на организацию рекламных кампаний. В связи с этим большой упор на этапе проектирования сайта организации делался на поисковую оптимизацию страниц сайта (SEO). Это подводило к мысли о том, чтобы и страницы поисковой системы сделать оптимизированными под поисковые запросы пользователей. Однако данный замысел не получилось осуществить в полной мере вследствие архитектурной ошибки, допущенной на этапе проектирования сервиса, и отсутствия учёта того фактора, что «поисковые пауки» не заполняют никаких форм [1, с. 245], тогда как ввод поискового запроса пользователем являлся обязательным условием для начала работ с сервисом.

Решением, учитывающим названный фактор, стала разработка нового сервиса, процесс которого вёлся параллельно с разработкой поисковика и занял не менее 10 месяцев работ и подключения 6 из 12 отделов организации. Этим сервисом стал обучающий курс (https://gold-race.org/remote-format), спроектированный по всем канонам поисковой оптимизации [1], т. е. содержащий такие страницы сайта, которые с большей вероятностью позволят пользователю найти ответ на свой вопрос, вводимый в поисковой системе Яндекса/Гугла/др.

ОБ ОГРАНИЧЕНИЯХ ИСКУСТВЕННОГО ИНТЕЛЛЕКТА В РАСПОЗНАВАНИИ НОВОГО ПОНЯТИЙНОГО АППАРАТА

Основная сложность, с которой мы столкнулись при стремлении полностью автоматизировать процесс переноса цитат, состояла в том, что цитаты из переносимых книг содержали новый понятийный аппарат, ранее малораспространённый даже в кругах философов, например, «программа жизни», «энергокомпозит души», «ячейка матрицы души», «отрицательная Система Бога» и др. Более полный их список можно посмотреть в [3]. В связи с этим построение интеллектуальных алгоритмов нейронных сетей, базирующихся на анализе семантической составляющей абзацев и агрегирующих на их основе заголовки цитат, представлялось процессом маловозможным, поскольку чем более сложным является поня-

тийный аппарат агрегируемых текстов, тем менее вероятно, как показал эмпирический опыт, можно встретить хорошо размеченные корпуса текстов для него, как, например, в (https://ling.hse.ru/krut).

ОБ АЛГОРИТМИЧЕСКОЙ СЛОЖНОСТИ ПОИСКОВОЙ СИСТЕМЫ

Разработанная нами поисковая система, на наш взгляд, не является алгоритмически сложным программным решением. Поскольку для поиска цитаты мы не использовали ни парсинга PDF-страниц посредством технологии Elasticsearch [4] или Solr, ни сложных методик формирования предложений для автозаполнений поисковых запросов, реализованных, например, в поисковой системе Google [1], ни оптимизаций производительности за счёт сокращения количества циклов [5, с. 294] на низкоуровневых языках программирования вроде C/C++, встраиваемых в веб через WebAssembly [6]. Функции поиска были выполнены на высокоуровневом языке программирования JavaScript посредством использования исключительно нативных методов данного языка, таких как Array#filter() (https://developer.mozilla.org/ru/docs/Web/JavaScript/Reference/Global_Objects/Array/filter) для фильтрации цитат по заголовку и основному тексту, Array#slice() (https://developer.mozilla.org/ru/docs/Web/JavaScript/Reference/Global_Objects/Array/slice) для построения пагинации и String#includes() (https://developer.mozilla.org/ru/docs/Web/JavaScript/Reference/Global_Objects/Array/slice) для поиска точного, регистронезависимого совпадения с запросом пользователя.

Основная сложность нашего проекта состояла, скорее в том, чтобы:

1. Технически содействовать в организации работы волонтёров и консультировать о технических возможностях и ограничениях реализации новых идей;
2. Управлять небольшой командой, состоящей из 2 разработчиков, 2 специалистов по заполнению JSON-модулей текстами цитат и 2 веб-дизайнеров. Учитывая, что разработчиков было всего 2 – первый (автор данной статьи) на тот момент имел 3 года опыта разработки клиентской части веб-приложений, а второй до перехода в проект писал на Ассем-

блере, имея опыт на JavaScript не более 1 года – требовалось мобилизовать силы и вести full-stack разработку, занимаясь видами работ сразу нескольких специалистов. В числе осваиваемых специальностей были фронтенд-разработка, бэкенд-разработка, DevOps и роль владельца продукта, принимавшего участие на собраниях членов Совета Центра «Золотая Раса» с последующим сбором требований и оповещением о технических ограничениях специалистов Отдела Информационных Технологий Центра «Золотая Раса».

ЗАКЛЮЧЕНИЕ

На июнь 2022 года в поисковую систему были добавлены текстовые модули по цитатам из 14 книг разного объёма представленности на сайте.

Благодарности

Выражаем благодарность Екатерине Вербовской (Германия), автору идеи поисковой системы по книгам Л.А. Секлитовой и Л.Л. Стрельниковой. Также благодарим М.А. Горкавенко (Москва) и В.Г. Шмакова (Московская обл.) за их помощь в организации волонтёрской работы по составлению цитат и Е.Е. Белковского (Беларусь) за разработку утилиты, проверяющей качество цитат на этапе их интеграции в виде текстовых модулей в кодовую базу (из расширений XLS/XLSX в JSON).

СПИСОК ЛИТЕРАТУРЫ

1. *Энж Э., Спенсер С. Стрикчиола Д.* SEO – искусство раскрутки сайтов. 3-е изд. СПб: 2017. 816 с.

2. *Ларман К.* Применение UML 2.0 и шаблонов проектирования, 3-е изд. : Пер. с англ. СПб.: ООО «Диалектика», 2020. 736 с. : ил. Парал. тит. англ.

3. Словарь космической философии. М.: Свет, 2021. 304 с. (Серия «За гранью непознанного»).

4. *Тарнбулл Д., Берримен Дж.* Релевантный поиск с использованием Elasticsearch и Solr. / пер. с англ. Киселев А. Н. М.: ДМК Пресс, 2018. 408 с.: ил.

5. *Таненбаум Э., Остин Т.* Архитектура компьютера. 6-е изд. СПб.: Питер, 2013. 816 с.: ил.

6. *Галлан Жерар*. WebAssembly в действии. СПб.: Питер, 2022. 496 с.: ил. (Серия «Библиотека программиста»).

DEVELOPING TECHNOLOGICAL CYCLE OF SEARCH SYSTEM THAT AGREGATES CITATIONS BY BOOKS

R. V. Mosolov[0000-0002-4399-4397]

Corporation "X5 Retail Group" (Moscow)

R.V.Mosolov@ya.ru

Abstract

In this article, we have described the technological cycle to develop the search system by 14 philosophical books by L.A. Seklitova, and L.L. Strelnikova. The cycle contained 6 steps of work. The ideas from the article may be useful to project, and develop a software, aggregating citations from books series, monographs, scientific periodicals, or scientific articles. For example, this experience may be useful for creating customized links on secondary sources that needs at a stage of writing scientific articles and design of presentations in Pedagogy. The search system is the result of 1 year work by the article author, and the group of around 30 volunteers. The system is represented a service, integrating in the web application. The technological stack contains Jade, CSS, JS, Node.js, Express.js, ESLint, Jest.

Keywords*: search system, searching system, search by books, search by book, search by citations, citations aggregator, aggregator by citations, books aggregate, citations data aggregators, develop search engine.*

REFERENCES

1. J*enzh Je., Spenser S. Strikchiola D.* SEO – iskusstvo raskrutki saĭtov. 3-e izd. SPb: 2017. 816 s.

2. *Larman K.* Primenenie UML 2.0 i shablonov proektirovanija, 3-e izd. : Per. s angl. SPb.: OOO «Dialektika», 2020. 736 s. : il. Paral. tit. angl.

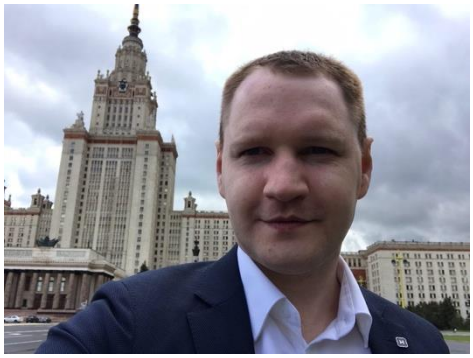
3. Slovar' kosmicheskoj filosofii. M.: Svet, 2021. 304 s. (Serija «Za gra-n'ju nepoznannogo»).

4. *Tarnbull D., Berrimen Dzh.* Relevantnyj poisk s ispol'zovaniem Elasticsearch i Solr. / per. s angl. Kiselev A. N. M.: DMK Press, 2018. 408 s.: il.

5. *Tanenbaum Je., Ostin T.* Arhitektura komp'jutera. 6-e izd. SPb.: Piter, 2013. 816 s.: il.

6. *Gallan Zherar.* WebAssembly v dejstvii. SPb.: Piter, 2022. 496 s.: il. (Serija «Biblioteka programmista»).

СВЕДЕНИЯ ОБ АВТОРЕ



МОСОЛО́В Роман Валерьевич – бакалавр социологии (Казанский федеральный университет (КФУ)), магистр компьютерных наук (Институт информационных технологий и интеллектуальных систем КФУ), старший разработчик в корпорации X5 Retail Group (Москва). В 2021 г. разрабатывал клиентскую часть сервисов изменения персональных данных и сбора данных (опросов) для 259 тыс. работников ТС «Пятёрочки», с мая 2022 г. по настоящее время участвует в разработке клиентской части сервиса базы знаний для 40 тыс. работников ТС «Перекрёстка».

Roman Valerievich MOSOLOV – Bachelor of Sociology (Kazan Federal University), Master of Computer Science (HS ITIS), senior developer at corporation X5 Retail Group (Moscow). Roman has developed client side of changing personal data and grabbing data (polls) services for 259K employees of “Pyaterochka” (supermarket chain) in 2021. At current time, he is developing client side of knowledge base service for 40K employees of “Perekrestok” (supermarket chain).

email: R.V.Mosolov@ya.ru

ORCID: 0000-0002-4399-4397

Материал поступил в редакцию 25 июля 2022 года