

Научная статья

УДК 004.550

DOI 10.25205/1818-7900-2021-19-3-40-49

Облачный сервис многомерной обработки количественных данных для решения геологических задач

Алексей Андреевич Загумёнов¹

Вера Викторовна Наумова²

Виталий Сергеевич Ерёмов³

¹ Институт автоматизации и процессов управления
Дальневосточного отделения Российской академии наук
Владивосток, Россия

^{2,3} Государственный геологический музей им. В. И. Вернадского
Российской академии наук
Москва, Россия

¹ truepikvic@gmail.com, <https://orcid.org/0000-0002-0501-5362>

² v.naumova@sgm.ru, <https://orcid.org/0000-0002-3001-1638>

³ vitaer@gmail.com, <https://orcid.org/0000-0002-5250-5743>

Аннотация

В работе описан разработанный облачный веб-сервис многомерной обработки количественных данных для решения широкого класса научных геологических задач. Вычислительный узел «Многомерные методы анализа данных» позволяет выполнять обработку табличных данных различными методами современного анализа данных с настройкой их параметров и визуализацией результатов. Узел включает в себя такие группы методов, как предобработка данных, описательная статистика, кластерный анализ, факторный анализ, корреляционный анализ, регрессионный анализ и др. Вычислительный узел «Многомерные методы анализа данных» входит в состав Вычислительно-аналитической геологической среды ГГМ РАН и интегрирован с ее сервисами. При этом вычислительный узел является самостоятельным облачным веб-сервисом, взаимодействие с которым происходит посредством REST API. Это позволяет обращаться к многомерным методам анализа данных, размещенным на вычислительном узле, широкому кругу пользователей, в том числе интегрировать его в другие информационные системы как стороннее приложение для обработки табличных данных.

Ключевые слова

облачные сервисы, геология, вычислительный узел, REST API

Благодарности

Работа выполняется в рамках Государственного задания ГГМ РАН по теме № 0140-2019-0005 «Разработка информационной среды интеграции данных естественнонаучных музеев и сервисов их обработки для наук о Земле»

Для цитирования

Загумёнов А. А., Наумова В. В., Ерёмов В. С. Облачный сервис многомерной обработки количественных данных для решения геологических задач // Вестник НГУ. Серия: Информационные технологии. 2021. Т. 19, № 3. С. 40–49. DOI 10.25205/1818-7900-2021-19-3-40-49

© Загумёнов А. А., Наумова В. В., Ерёмов В. С., 2021

ISSN 1818-7900 (Print). ISSN 2410-0420 (Online)

Вестник НГУ. Серия: Информационные технологии. 2021. Том 19, № 3. С. 40–49

Vestnik NSU. Series: Information Technologies, 2021, vol. 19, no. 3, pp. 40–49

Cloud Service for Multidimensional Processing of Quantitative Data for Solving Geological

Aleksei A. Zagumennov¹, Vera V. Naumova²
Vitaliy S. Eremenko³

¹ Institute of Automation and Control Processes
of the Far East Branch of the Russian Academy of Sciences

^{2,3} Vladivostok, Russian Federation

² V. I. Vernadsky State Geological Museum of the Russian Academy of Sciences
Moscow, Russian Federation

¹ truepikvic@gmail.com, <https://orcid.org/0000-0002-0501-5362>

² v.naumova@sgm.ru, <https://orcid.org/0000-0002-3001-1638>

³ vitaer@gmail.com, <https://orcid.org/0000-0002-5250-5743>

Abstract

The study describes the developed cloud web service for multidimensional processing of quantitative data for solving a wide class of scientific geological tasks. The computing node “Multidimensional methods of data analysis” provides processing of tabular data using various methods of modern data analysis and allows to set their parameters and visualize the results. The node includes wide range of methods such as data preprocessing, descriptive statistics, cluster analysis, factor analysis, correlation analysis, regression analysis. Computing node “Multidimensional methods of data analysis” is a part of Computational analytical geological environment of State Geological Museum of RAS and is integrated with its services. At the same time, the computing node is an independent cloud web service which implements REST API for interaction with it. This allows a wide range of users to access multidimensional data analysis methods located on a computing node and provides capabilities of its integration into information systems as a third-party application for processing tabular data.

Keywords

cloud services, geology, computing node, REST API

Acknowledgements

The study is supported by the Government contract no. 0140-2019-0005 “Development of an information environment for integrating data from natural science museums and services for their processing for Earth sciences”

For citation

Zagumennov A. A., Naumova V. V., Eremenko V. S. Cloud Service for Multidimensional Processing of Quantitative Data for Solving Geological. *Vestnik NSU. Series: Information Technologies*, 2021, vol. 19, no. 3, p. 40–49. (in Russ.) DOI 10.25205/1818-7900-2021-19-3-40-49

Введение

При геологических исследованиях быстрыми темпами накапливается большое количество геологической информации: результаты геологической документации буровых скважин, горных выработок и естественных обнажений, спектральных и химических анализов руд, горных пород и минералов, данные геофизических и геохимических измерений и др. Одно из важнейших направлений научно-технического прогресса в геологии состоит в широком внедрении автоматизированных методов накопления, хранения, обработки и передачи геологической информации с целью повышения эффективности геологических исследований.

Сегодня для решения целого ряда геологических задач используются многомерные методы анализа данных. Данные методы применяются с целью обнаружения скрытых закономерностей, выявления наиболее существенных связей между переменными путем анализа взаимосвязи большого количества этих переменных. К таким методам относят факторный, кластерный, дисперсионный, регрессионный анализ, многомерное шкалирование, анализ канонических переменных и др.

На рынке представлено немало традиционных профессиональных пакетов, обеспечивающих эти функции для анализа геологических данных: SPSS, STATISTICA, MATLAB и др.

Среди современных веб-инструментов анализа данных (в том числе и Big Data) необходимо выделить нижеследующие.

Orange – проект, содержащий набор инструментов для визуализации, обработки и анализа данных. Данный пакет инструментов поможет создать интерактивный рабочий процесс для анализа и визуализации данных с помощью точечных диаграмм, гистограмм, деревьев, дендрограмм, сетей и тепловых карт.

NodeXL – программное обеспечение для анализа данных, визуализации зависимостей и сетей. Вот некоторые его возможности:

- визуализация собственной сети в соответствии с требованиями;
- анализ социальных сетей и определение ключевых подписчиков;
- анализ контента позволяет глубоко анализировать все слова, хэштеги и URL-адреса, встроенные в сообщения, твиты и т. д.;
- импорт наборов данных из различных источников;
- автоматизация и отчеты.

Salesforce reports & dashboards – профессиональный инструмент для анализа и визуализации больших наборов данных. Вот некоторые его особенности:

- фильтрация, суммирование, создание формул;
- визуализация данных с обновлением в режиме реального времени;
- управление данными из административной части CRM-системы.

Стоит отметить, что Salesforce не относится к бесплатным инструментам.

RapidMiner – еще один из современных инструментов. Комплекс включает в себя IDE собственной разработки и сервер для хранения выполненных процессов анализа, общего доступа к информации аналитиками и подготовки результатов для выдачи в веб-интерфейсе. Платформа позволяет быстро оперировать данными из разных источников, запускать проверки моделей машинного обучения, и всё это в одном месте с одним набором программного обеспечения.

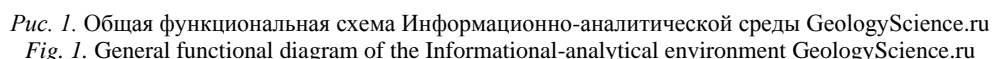
Однако при решении геологических задач не всегда требуется полная функциональность профессиональных пакетов, а необходимы лишь отдельные методы обработки данных.

Основная цель работ – обеспечение геологов облачным сервисом, в единой точке доступа предоставляющим методы многомерной обработки количественных данных для решения геологических задач на основе открытых решений, для более эффективного использования, анализа распределенных геологических ресурсов и возможности построения прогнозов на основе геологических моделей.

Для этого был разработан Вычислительный узел «Многомерные методы анализа данных» в составе Вычислительно-аналитической геологической среды ГГМ РАН (далее Вычислительная среда).

Вычислительно-аналитическая геологическая среда ГГМ РАН

С 2017 г. в Государственном геологическом музее им В. И. Вернадского РАН ведется разработка и адаптация методов и технологий обработки и анализа территориально распределенной разнотипной геологической информации и сервисов ее обработки. На основе созданных подходов, разработанных методов и технологий реализуется информационно-аналитическая среда для поддержки и сопровождения научных исследований в геологии, осуществляющая интеграцию территориально распределенной геологической информации с использованием специализированных служб ее анализа и обработки (<http://geologyscience.ru>) (рис. 1) [1]. Авторы предполагают, что разработанная программная платформа управления тематическими сервисами обработки и анализа, которая является частью информационно-аналитической среды, обеспечит пользователям доступ к хранилищам современных наукоемких алгоритмов и вычислительных ресурсов, необходимых для оперативной обработки больших массивов геологических данных.



В результате проведенных работ авторами разработана цифровая платформа по обработке и анализу геологических данных DPAР [3]. Платформа представляет собой программный продукт на языке программирования Java в формате веб-приложения с использованием фреймворка Spring Boot. Она реализована с использованием принципов слабосвязанной архитектуры. Для взаимодействия с пользователем реализован веб-интерфейс с JavaScript фреймворком Angularjs. При этом разработан программный интерфейс с использованием архитектурного стиля REST для взаимодействия с платформой в формате «система – система».

Серверная технология DPAR обеспечивает безопасное подключение к Вычислительной среде совместной работы, которая превращает пользовательские разрозненные данные в безопасный для анализа фонд информации. Масштабируемая технология создания каталогов и управления большими наборами данных позволяет администрировать одновременный доступ с разными пользовательскими настройками и допусками безопасности. Платформа позволяет пользователям эффективно каталогизировать, управлять, визуализировать большие пространственные данные, что поможет им, минимизировать риск и принимать более обоснованные решения, увеличивающие возможность успеха обнаружения месторождений.

Использование цифровой платформы по обработке и анализу геологических данных позволяет в дальнейшем расширять возможности разработчиков информационно-аналитических сред для поддержки и сопровождения научных исследований.

В Вычислительной среде реализован доступ к следующим вычислительным узлам и платформам.

- **Вычислительный узел «Многомерные методы анализа данных»**, разрабатываемый в ГИМ РАН, который позволяет выполнять обработку табличных данных различными методами анализа данных с настройкой их параметров и визуализацией результатов. Вычислительный узел включает в себя такие группы методов, как предобработка данных, описательная статистика, кластерный, факторный, корреляционный, регрессионный анализ и др.

- **Обработка петролого-геохимических данных.** В Институте физики Земли РАН разработана интерактивная база методов обработки петролого-геохимических данных [4]. Данная система предоставляет сервисы построения спайдерграмм, гистограмм и классификационных диаграмм; сервис идентификации минералов по их химическому составу; сервис интерпретации состава минерала и разложение на минералы и т. д. Интерфейс взаимодействия с сервисами построен на основе REST архитектуры.

- **Структурный анализ публикаций.** В междисциплинарном центре математического и вычислительного моделирования (Университет Варшавы, Польша) разработан сервис для извлечения метаданных из научных публикаций [5]. Метаданные включают в себя авторов, аффилиацию, аннотацию, ключевые слова, название журнала, объем, год выпуска, разобранные библиографические ссылки, структуру разделов документа, заголовки разделов и абзацы. Интерфейс взаимодействия с сервисами построен на основе REST архитектуры.

- **Обработка естественного языка.** В Университете Шеффилда в рамках проекта GATE (General Architecture for Text Engineering) разработан ряд сервисов по обработке текстовых данных для различных языков [6]. Для обработки текстовых данных на русском языке предоставляются сервисы по определению частей речи слов, а также выделению именованных сущностей, таких как имена и фамилии, названия организаций, географические названия, даты, денежные единицы и т. д. Интерфейс взаимодействия с сервисами построен на основе REST архитектуры.

В Вычислительную среду также включены следующие облачные сервисы. Одним из наиболее популярных программных продуктов обработки табличных данных в интерактивном режиме является **Excel** из пакета Microsoft Office. Данный программный продукт содержит ряд инструментов для редактирования данных, построения различных диаграмм, использования встроенных процедур анализа и создания собственных. Компания «Microsoft» разработала бесплатную облачную версию продукта Excel (<https://www.office.com/>). Зарегистрировавшись, можно использовать полноценную веб-версию для данных, расположенных в облачном хранилище Microsoft Onedrive.

Для анализа пространственных данных наиболее подходящим инструментом являются геоинформационные системы (ГИС), позволяющие в интерактивном режиме взаимодействовать с объектами на карте, применяя к отдельным слоям и объектам доступные средства анализа. Наиболее подходящим для геологии облачным решением является **ArcGIS Online**, разрабатываемый компанией ESRI. Portal for ArcGIS – это инфраструктура ArcGIS Online, функционирующая в защищенной ИТ-среде или в частном облаке организации (под контро-

лем сетевого экрана или в полностью изолированной локальной сети). Портал позволяет создавать карты, каталогизировать и анализировать пространственные данные с помощью удобного, интуитивного понятного интерфейса.

Для анализа спутниковых данных одним из лидеров среди облачных сервисов является **An Open Source Geospatial Data Management & Analysis Platform**. Open Data Cube (ODC) – это некоммерческий проект с открытым исходным кодом, который был создан для обеспечения доступа, управления и анализа больших объемов данных ГИС по мониторингу Земли. Он представляет собой общий аналитический фреймворк, содержащий наборы структурированных данных и инструментов, которые позволяют проводить анализ больших коллекций пространственных данных. ODC был разработан для анализа пространственных данных на больших временных промежутках, однако его можно использовать на любых наборах пространственных данных. Данные могут представлять собой модели высот, геофизические сетки, интерполированные поверхности и т. д. Ключевой особенностью ODC является возможность сохранения каждого уникального элемента набора пространственных данных, в отличие от многих других методов работы с большими коллекциями пространственных данных.

Вычислительный узел «Многомерные методы анализа данных»

Вычислительный узел позволяет выполнять обработку табличных данных различными методами анализа данных с настройкой их параметров и визуализацией результатов. Вычислительный узел включает в себя такие группы методов, как предобработка данных, описательная статистика, кластерный, факторный, корреляционный, регрессионный анализ и др. (рис. 2) (<https://service.geologyscience.ru/service?nodeId=1>). Перечень предлагаемых пользователю методов постоянно пополняется.

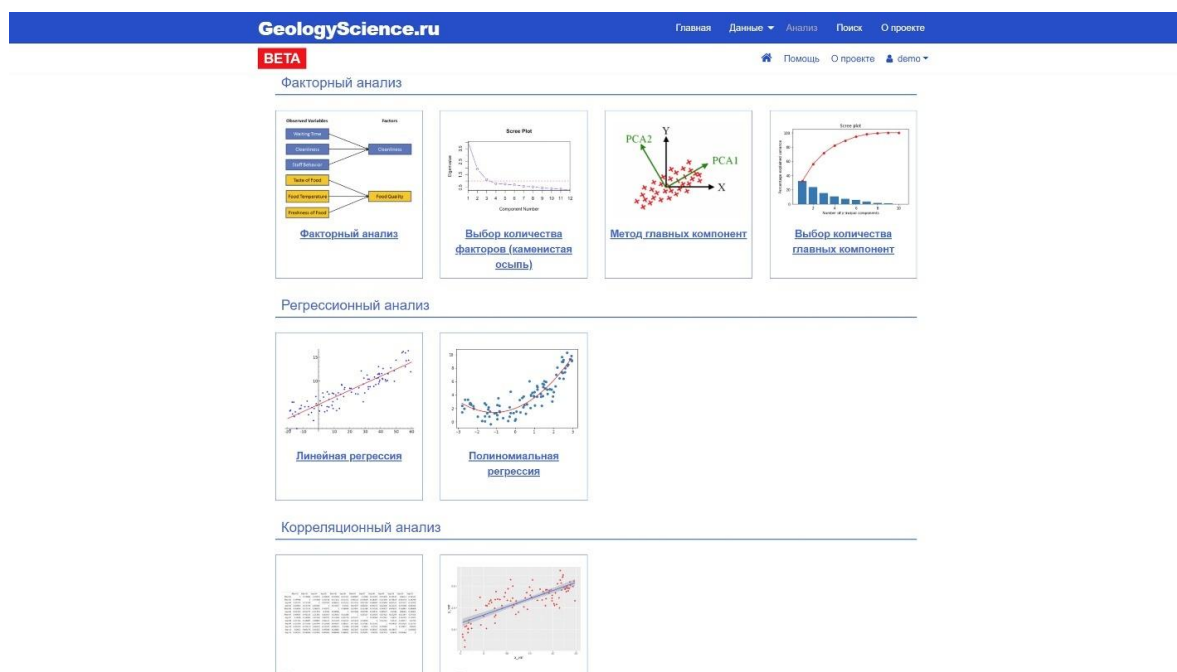


Рис. 2. Страница выбора многомерных методов анализа данных

Fig. 2. Multidimensional methods of data analysis selection webpage

Выбор конкретных методов для включения в вычислительный узел обуславливается ориентированностью узла в первую очередь на решение геологических задач. Работа с геологическими данными часто опирается на статистические методы анализа данных, проверку гипотез, поиск зависимостей и корреляций. Результаты применения этих методов позволяют сделать выводы об исследуемых данных и принять решения о применении других методов: кластеризации, классификации и др. При этом важно иметь возможность настройки параметров методов для более детального анализа геологических данных.

Вычислительный узел состоит из следующих компонент (рис. 3):

- веб-сервис для получения запросов от внешних пользователей или систем на обработку данных выбранным методом с указанными параметрами, проверку статуса обработки и получения результатов работы метода;
- реестр методов, содержащий перечень методов, их параметров, а также непосредственную реализацию методов в виде скриптов на языке Python;
- очередь задач, в которую помещаются поступающие запросы на обработку данных, имеющую возможность отслеживания статуса выполнения задач;
- исполнители задач, которые при получении новой задачи из очереди обращаются к реестру методов и выполняют соответствующий скрипт;
- хранилище временных файлов, в котором для каждой задачи хранятся входные файлы и файлы результатов.

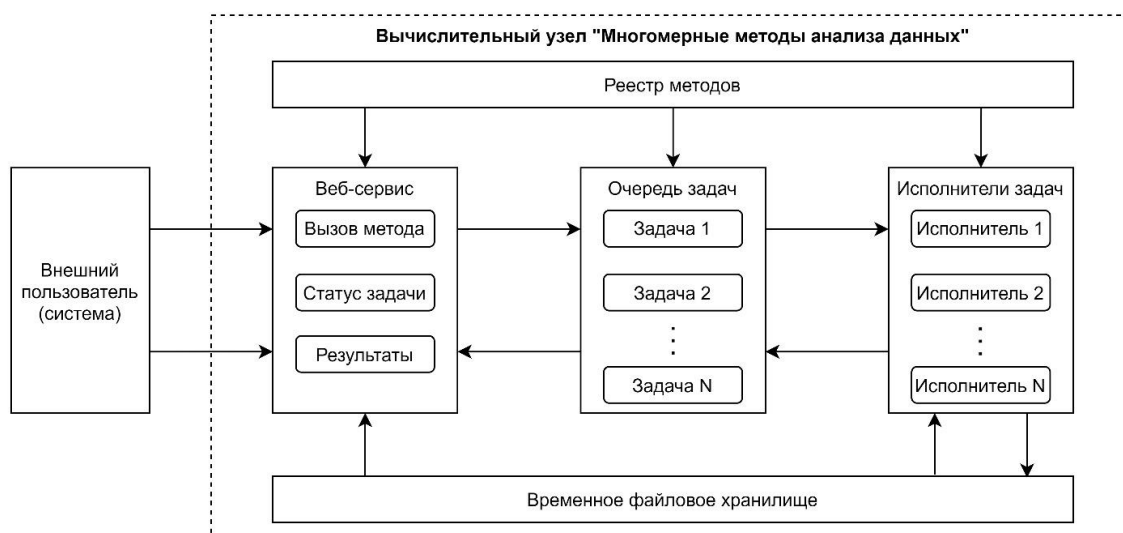


Рис. 3. Принципиальная схема Вычислительного узла «Многомерные методы анализа данных»

Fig. 3. Schematic diagram of the Computing node "Multidimensional methods of data analysis"

Вычисления производятся в среде Python с использованием известных открытых пакетов обработки данных: Scikit-learn, Pandas, Matplotlib и др. Обработка поступающих запросов на обработку данных производится фреймворком FastAPI посредством REST API, реализованного по стандарту OpenAPI (<http://spec.openapis.org/oas/v3.0.3>), с использованием очереди вычислительных задач, реализованной на основе NoSQL базы данных Redis.

При этом отдельные компоненты системы (веб-сервис, база данных с очередью задач, исполнители задач) разворачиваются с применением контейнеров Docker. Такая архитектура позволяет разделить обработку запросов и тяжелые вычисления больших объемов данных, что обеспечивает отказоустойчивость и масштабируемость узла.

Вычислительный узел «Многомерные методы анализа данных» является самостоятельным облачным веб-сервисом (<https://analysis.geologyscience.ru>), что позволяет использовать его отдельно, посредством REST API, реализованного по стандарту OpenAPI 3.0. Это обеспечивает возможность обращаться к многомерным методам анализа данных, размещенным на вычислительном узле, широкому кругу пользователей, в том числе интегрировать его в другие информационные системы как стороннее приложение для обработки табличных данных. Кроме того, использование стандарта OpenAPI делает узел самодокументированным, а при изменении списка методов документация обновляется автоматически (<https://analysis.geologyscience.ru/docs>). Согласно классификации облачных сервисов, приведенных в работе [7], вычислительный узел «Многомерные методы анализа данных» реализует модель «программное обеспечение как сервис» (Software as a Service, SaaS), т. е. пользователи сервиса получают доступ к возможностям сервиса посредством облачных технологий, без необходимости устанавливать и конфигурировать программное обеспечение самостоятельно.

Взаимодействие с Вычислительно-аналитической средой по обработке и анализу геологических данных

Взаимодействие с платформой по обработке и анализу геологических данных осуществляется с использованием разработанного REST интерфейса. В Вычислительной среде реализован модуль взаимодействия с внешними веб-сервисами (WSIM), предоставляющий единый интерфейс доступа к внешним сервисам обработки и анализа, работающим по принципу «запрос – ответ». В качестве интерфейса доступа используется OGC Web Processing Service 1.0. Каждый используемый Вычислительной средой метод публикуется в виде отдельного процесса внутри WPS сервиса. Для реализации WPS сервиса используется программный пакет с открытым исходным кодом GeoServer. GeoServer содержит встроенные компоненты для реализации очереди заданий, разграничения доступа к различным процессам и возможность как синхронного, так и асинхронного выполнения процессов. Для реализации модуля используются программные библиотеки GeoTools и 52 North WPS. Взаимодействие платформы с данным модулем осуществляется через WPS интерфейс.

В Вычислительной среде реализован доступ к общему облачному хранилищу пользователя, используемому в информационно-аналитической системе GeologyScience.ru, при помощи отдельного программного модуля (MICS). Хранилище позволяет размещать пользовательские данные для их дальнейшего использования в других подсистемах, в том числе для обработки и анализа. Результаты обработки также могут быть сохранены пользователем в облачное хранилище на длительный срок для их последующего использования в других сервисах. Взаимодействие с платформой осуществляется посредством REST интерфейса.

Для проведения предварительного анализа количественных данных в Вычислительной среде используется программный продукт Excel с доступом через интерактивную облачную платформу Microsoft Office Online, описанный ранее.

Все используемые в Вычислительной среде сервисы, включая сервисы вычислительного узла «Многомерные методы анализа данных», публикуются в каталоге сервисов среды. Основные функции каталога: обеспечение пользователя информацией об имеющихся в Вычислительной среде сервисах с возможностью получения информации о каждом зарегистрированном сервисе, включая общее описание, данных о поставщике сервиса и технической информации для организации взаимодействия. Информация из данного модуля используется в модуле мониторинга для отслеживания состояния каждого внешнего сервиса, зарегистрированного в Вычислительной среде. Взаимодействие с платформой осуществляется с использованием REST интерфейса.

Дальнейшие перспективы развития Вычислительного узла

Дальнейшее развитие Вычислительного узла предполагается вести в двух контекстах: прикладном и системном. В рамках прикладного контекста планируется расширение количества предлагаемых методов анализа данных, добавление пространственных методов обработки геологических данных, реализация механизма цепочек обработки геологических данных. Развитие узла в системном контексте предполагает реализацию возможности двусторонней работы с узлом обработки (по механизму Webhook или протоколу WebSocket), создание средств авторизации и учета пользователей узла, разработку модуля аналитики поведения пользователей, расширение возможностей интеграции вычислительного узла в другие информационные системы.

Список литературы

1. **Наумова В. В., Платонов К. А., Еременко В. С., Патук М. И., Дьяков С. Е.** Информационно-аналитическая среда для поддержки научных исследований в геологии: текущее состояние и перспективы развития // Тр. XVII Междунар. конф. «Распределенные информационно-вычислительные ресурсы. Цифровые двойники и большие данные. (DICR-2019)». Новосибирск, 2019. С. 139–147. DOI 10.25743/ICT.2019.70.61.021
2. **Eremenko V. S., Naumova V. V., Platonov K. A., Dyakov S. E., Eremenko A. S.** The main components of a distributed computational and analytical environment for the scientific study of geological systems. *Russian Journal of Earth Sciences*, 2018, vol. 18, iss. 6. DOI 10.2205/2018ES000636
3. **Еременко В. С., Наумова В. В., Загумёнов А. А., Булов С. В.** Облачные технологии для развития территориально распределенной вычислительно-аналитической геологической среды // Вычислительные технологии. 2021. Т. 26, № 1. С. 86–98.
4. **Иванов С. Д.** Интерактивный реестр геосенсоров на основе веб-приложения // Компьютерные исследования и моделирование. 2016. Т. 8, № 4. С. 621–632.
5. **Tkaczyk D., Szostek P., Fedoryszak M., Dendek P., Bolikowski L.** CERMINE: automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition*, 2015, vol. 18, no. 4, pp. 317–335.
6. **Maynard D., Bontcheva K., Augenstein I.** Natural Language Processing for the Semantic Web. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 2016, vol. 6, no. 2, pp. 1–194.
7. **Moser L., Thuraisingham B., Zhang J.** Services in the cloud. *IEEE transactions on services computing*, 2015, vol. 8, no. 2, pp. 172–174.

References

1. **Naumova V. V., Platonov K. A., Eremenko V. S., Patuk M. I., Dyakov S. E.** Information and analytical environment for supporting scientific research in geology: current state and development prospects. In: Proceedings of the XVII International conference “Distributed information and computing resources. (DICR-2019)”. Novosibirsk, 2019, pp. 139–147. (in Russ.) DOI 10.25743/ICT.2019.70.61.021
2. **Eremenko V. S., Naumova V. V., Platonov K. A., Dyakov S. E., Eremenko A. S.** The main components of a distributed computational and analytical environment for the scientific study of geological systems. *Russian Journal of Earth Sciences*, 2018, vol. 18, iss. 6. DOI 10.2205/2018ES000636
3. **Eremenko V. S., Naumova V. V., Zagumennov A. A., Bulov S. V.** Cloud technologies for development of geographically distributed computational and analytical Geological environment. *Computational Technologies*, 2021, vol. 26, no. 1, pp. 86–98. (in Russ.)

4. **Ivanov S. D.** Interactive Web Application Based Geosensors Registry. *Computer research and modeling*, 2016, vol. 8, no. 4, pp. 621–632. (in Russ.)
5. **Tkaczyk D., Szostek P., Fedoryszak M., Dendek P., Bolikowski L.** CERMINE: automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition*, 2015, vol. 18, no. 4, pp. 317–335.
6. **Maynard D., Bontcheva K., Augenstein I.** Natural Language Processing for the Semantic Web. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 2016, vol. 6, no. 2, pp. 1–194.
7. **Moser L., Thuraisingham B., Zhang J.** Services in the cloud. *IEEE transactions on services computing*, 2015, vol. 8, no. 2, pp. 172–174.

Информация об авторах

Алексей Андреевич Загумённый, научный сотрудник

Researcher ID R-4407-2016

Вера Викторовна Наумова, доктор геолого-минералогических наук

Researcher ID J-9039-2018

Виталий Сергеевич Ерёмченко, младший научный сотрудник

Researcher ID Q-3678-2016

Information about the Authors

Aleksei A. Zagumennov, Researcher

Researcher ID R-4407-2016

Vera V. Naumova, Doctor of Sciences (Geology and Mineralogy)

Researcher ID J-9039-2018

Vitaliy S. Eremenko, Junior Researcher

Researcher ID Q-3678-2016

Статья поступила в редакцию 20.06.2021;

одобрена после рецензирования 01.08.2021; принята к публикации 01.08.2021

The article was submitted 20.06.2021;

approved after reviewing 01.08.2021; accepted for publication 01.08.2021