ИНФОРМАЦИОННО-АНАЛИТИЧЕСКАЯ СРЕДА ДЛЯ ПОДДЕРЖКИ НАУЧНЫХ ИССЛЕДОВАНИЙ В ГЕОЛОГИИ: ТЕКУЩЕЕ СОСТОЯНИЕ И ПЕРСПЕКТИВЫ РАЗВИТИЯ

-

Наумова В.В. [(1)] *, Платонов К.А.* [(1)]*, Еременко В.С.* [(1)]*, Патук М.И.* [(1)] *, Дьяков С.Е.*[(2)]

[(1)] ФГБУН Государственный геологический музей им. В.И. Вернадского РАН, г. Москва
[(2)] ФГБУН Институт автоматики и процессов управления ДВО РАН, г. Владивосток

DOI: 10.25743/ICT.2019.70.61.021

В работе описана 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 разработка базовой основы информационно-аналитической среды для поддержки и сопровождения научных исследований в геологии, осуществляющей интеграцию территориально распределенной геологической информации с использованием специализированных служб её анализа и обработки. Авторы предполагают, что разработанная 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.

Ключевые слова. Информационно-аналитическая среда, интеграция разнотипной территориально распределенной геологической информации, вычислительно-аналитическая обработка геологической информации.

INFORMATION AND ANALYTICAL ENVIRONMENT TO SUPPORT SCIENTIFIC RESEARCH IN GEOLOGY: CURRENT STATUS AND PROSPECTS FOR DEVELOPMENT

Naumova V.V. [(1)]*, Platonov K.A.* [(1)]*, Eremenko V.S.* [(1)]*, Patuk M.I.* [(1)]*, Dyakov S.E.* [(2)]

[(1)]V.I. Vernadsky State geological museum of RAS, Moscow
[(2)]Institute of automation and control processes of FEB RAS, Vladivostok

Development and adaptation of methods and technologies for processing and the analysis of territorially distributed polytypic geological information and services of its processing is described in this paper. On the basis of the created approaches, the developed methods and technologies, the carried-out design the basis of the information - analytical environment for support and maintenance of scientific research in geology which is carrying out integration of the polytypic territorially distributed geological information with use of specialized services of its analysis and processing is realized. Authors suppose that the developed platform of management of thematic services of processing and the analysis which is a part of the information - analytical environment will provide to users access to the storages of the modern knowledge-intensive algorithms and computing resources necessary for expeditious processing of larger arrays of polytypic geological data. The environment is intended for support and maintenance of scientific researches in geology.

Information-analytical environment, integration of heterogeneous geographically distributed geological information, computational-analytical environment of geological information processing.

Введение. Интеграция информационных и вычислительных ресурсов в единую среду и организация доступа к ним является одним из важнейших направлений развития современ-

ных информационных технологий. [5; 6].

Системы открытого доступа к данным и системам обработки. Термин «открытый доступ» впервые был упомянут на Будапештской конференции по открытому доступу в феврале 2002 г. С тех пор его смысл практически не изменился: Термин определяется как бесплатный, оперативный, постоянный, полнотекстовый, онлайновый доступ к научной информации. В ходе 69-й Генеральной конференции ИФЛА на семинаре «Информационные технологии и работа группы метаданных Dublin Core» были сформулированы принципы, на которых базируется идеология «Открытого архива»: консолидация в мировом масштабе архивов научных материалов; свободный доступ к архивам (к метаданным); согласованные интерфейсы архивов и поставщиков информации; простота использования; применение существующих стандартов – HTTP, XML, Dublin Core, MARC, MARCXML.

В настоящее время для интенсификации научных исследований и развития научных коммуникаций разрабатываются системы открытого доступа к научным публикациям, к архивам научной информации, к информационным системам, к данным естественнонаучных музеев и др.

Актуальной задачей является обеспечение открытого доступа к информации и цифровым данным, а также к системам анализа и обработки для наук о Земле.

В настоящее время наиболее известными являются следующие системы:

Digital Earth Australia (http://www.ga.gov.au/dea/home) - это реализация правительством Австралии платформы анализа с открытым исходным кодом, разработанной в рамках инициативы Open Data Cube (ODC). Программа DEA предоставляет код, документацию, руководства по использованию, учебные пособия и поддержку для международных пользователей Open Data Cube. Открытый куб данных (ODC) является глобальной инициативой по увеличению возможностей использования спутниковых данных, предоставляя пользователям доступ к свободным и открытым технологиям управления данными и платформам анализа. Применение бесплатных и открытых спутниковых данных для экологических, экономических и социальных задач может предоставить информацию и приложения, которые оказывают большое влияние на местные, региональные и глобальные масштабы. Достижения в облачных вычислениях и наличие свободных и открытых технологий, таких как Open Data Cube, означают, что различные страны без локальной инфраструктуры для обработки больших объемов данных могут получить доступ к данным и вычислительным мощностям для создания соответствующих приложений и информирования о принятии решений.

Основная цель *U.S. Geoscience Information Network* (http://usgin.org) облегчить открытый доступ к совместимым цифровым данным и программному обеспечению в науках о Земле. USGIN стандарты, протоколы и задачи - наследие National Geothermal Data System (NGDS), системы совместного использования данных, которая обеспечивает доступ к информации о геотермических ресурсах.

Британская геологическая служба имеет широкий спектр наборов данных и постоянно расширяет доступ к ним к ним, публикуя большое количество данных на портале *OpenGeoscience BGS* (http://www.bgs.ac.uk/opengeoscience). Услуги, доступные в OpenGeoscience включают: просмотр геологических данных через поисковое окно геологической карты Великобритании, а также используя WMS; доступ к более чем миллиону ска-

нирований фотографий геологических разрезов и скважин, а также фотографиям из геологического фотоархива GeoScienic; просмотр опубликованных бумажных карт с 1832 по 2014 год и публикаций с 1835 года по настоящее время.

Портал EarthChem, поддерживаемый Колумбийским университетом (http://www.earthchem.org, содержит более 860 тыс. образцов из 20 тыс. научных геологических публикаций и предоставляет возможность анализа и визуализации на карте содержимого геохимических баз данных таких, как GeoRock, PetDB, CedDB и др.

Проект Государственного геологического музея им В.И.Вернадского РАН (ГГМ РАН) «Разработка Информационно-аналитической геологической среды для поддержки научных исследований GeologyScience.ru»

В 2014-2017 г. авторами проводились работы по проектированию и реализации первой версии Интернет - инфраструктуры для поддержки и сопровождения научных геологических исследований на Дальнем Востоке России [3; 4].

Основная цель данного Проекта заключается в организации единой очки доступа к геологическим данным на территорию России и системам их обработки с использованием возможностей поиска данных в территориально распределенных разнородных источниках, а также с использованием территориально-распределенных вычислительно-аналитических узлов для обработки данных, взаимодействие с которыми осуществляется с использованием технологии web-сервисов. Интеграция разнотипных геологических данных и сервисов обработки в единую информационно-аналитическую среду на основе единых политик обеспечивает возможность комплексного анализа информации и позволит получать качественно новые знания о геологических объектах.

В основе предложенного подхода лежит слабосвязанная блочная инфраструктура, основанная на различии в типах геологических данных: пространственных, количественных, библиографических и основанных на экспертных знаниях. В каждом отдельном информационном блоке Среды для интеграции, хранения и поиска данных применяются различные подходы и технологические решения.

Сформулированы основные требования для организации Среды:

- Доступ к информационным ресурсам на основе международных стандартов и единых политиках;
- Сквозной поиск информации в Среде как на логическом, так и на физическом уровне;
- Организация мониторинга территориально распределенных источников данных и вычислительных узлов, а также основных узлов Среды;
- Поддержка сквозной авторизации и разграничения прав.

На рис.1 представлена обобщенная схема Информационно-аналитической геологической среды. Среда содержит 2 основных уровня: информационный и вычислительный (рис.1).

Информационный уровень Среды. Объединение тематических ресурсов в общую интегрированную информационную инфраструктуру поддержки научных исследований позволяет получить прямой доступ ко всем узлам Системы, уменьшает количество действий пользователя при общении с каждым узлом, позволяет отдельным узлам и отдельным сервисам данных узлов быть интегрированным в другие информационные системы, облегчает процесс администрирования Системы.

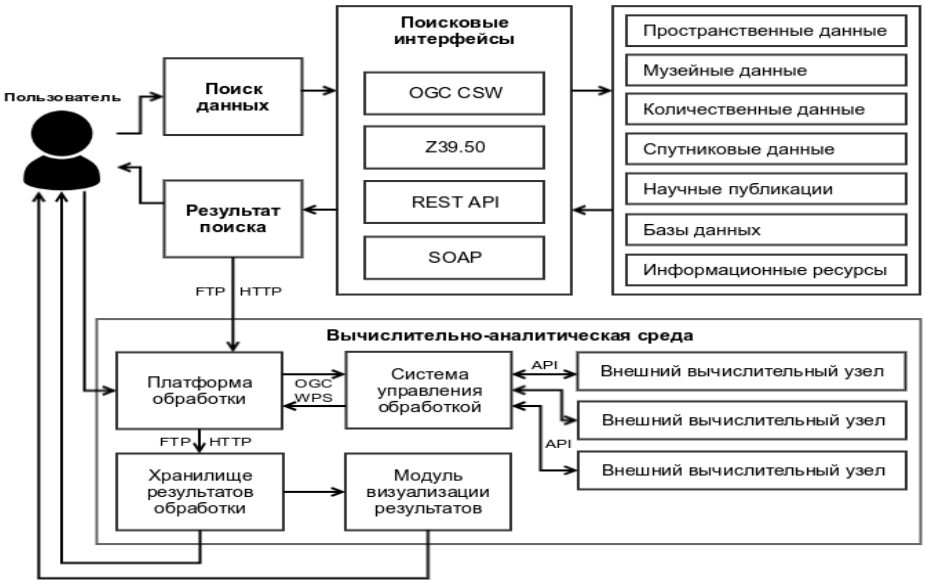

Рис. 1 Обобщенная функциональная схема Информационно-аналитической геологической среды

Интеграция информационных ресурсов подразумевает исполнение функций [6], обеспечивающих:

1. Доступ ко всем интегрированным ресурсам через единые пользовательские интерфейсы по единым протоколам;
2. Сквозной поиск во всем множестве интегрированных информационных ресурсов, а также в их логических и физических подмножествах;
3. Извлечение информации в единых форматах;
4. Управление ресурсами и доступом к ним в соответствии с едиными политиками;
5. Контроль целостности и доступности сервисов для всех ресурсов;
6. Сбор статистической информации об использовании ресурсов.

Источники информации - территориально распределенные Интернет-ресурсы, информация в которых основана на стандартизованных метаданных, и программные решения которых допускают применение стандартизованных протоколов для ее автоматической интеграции в создаваемую инфраструктуру, а также научные материалы научных организаций, библиотек, центров данных и др.

Нами сформулированы основные требования к информационным блокам Среды:

- Однородность данных в рамках блока;
- Извлечение информации из территориально –распределенных источников должно быть осуществлено в единых форматах для каждого блока;
- Наличие БД метаданных в стандартах данного типа данных;
- Наличие API поиска.

Блок «Научные публикации» представляет собой репозиторий открытого доступа, созданный на основе DSpace, 6.3. Основой информации являются научные статьи, монографии, диссертации, авторефераты диссертаций, тезисы докладов, материалы конференций.

Для поиска и извлечения информации из других репозиториев создан скрипт на языке PHP. Извлекаемая информация фильтруется, т.е. автоматически анализируется на совпадение со словарем геологических терминов. Словарь создан на основе ключевых слов ~ 2000 публикаций по тематике репозитория. Оптимальным является наличие 3-х совпадений со

словарем. При этом удается выбрать около 90 % источников, соответствующих тематике репозитория. Оставшиеся 10 % подвергаются ручной обработке. Эта информация служит основой для корректировки словаря.

Большое количество информации в открытом доступе представляют собой тексты в формате PDF. Для извлечения метаданных из таких публикаций используется свободно распространяемое программное обеспечение: Cermine- Content Extractor and Miner, FPDI – коллекция PHP классов для обработки PDF документов , PDFMiner – программное обеспечение на Python для извлечения текстовой информации из PDF.

Извлекаемая из других репозиториев и из файлов PDF информация преобразуется в формат SIP, доступный для импорта в DSpace стандартными средствами. Для улучшения поиска информации в репозитории к существующим стандартным в DSpace поисковым тегам был добавлен тег УДК (универсальная десятичная классификация). Данная информация извлекается в полуавтоматическом режиме из выгруженного из DSpace бэкапа в формате SIP в текстовый файл с последующей загрузкой SQL скриптом в таблицу PostgreSQL DSpace.

Блок количественных данных. Разрабатываемый блок является единой точкой доступа к территориально-распределенной количественной информации по геологии через унифицированный интерфейс [10]. Данные интегрируются на логическом уровне, с использованием глобальной схемы метаданных DataCite. В качестве источников данных выбраны научные публикаций, авторские базы данных, мировые Центры данных, а также пользовательские таблицы экспериментальных данных. Обнаружение и получение данных в информационной системе возможно по протоколам OAI и RESTfull - интерфейсу. Интегрируемая информация, если необходимо, подвергается трансформации авторскими алгоритмами в формат электронных таблиц и генерируются метаданные.

Для обеспечения пользовательского доступа к геологическим картам России разработан **Блок доступа к пространственным данным.** Доступ реализован с использованием технологии каталогизации метаданных пространственных данных на основе международного стандарта сервиса каталога OGC Catalogue Service for the Web (OGC CSW). Сервис каталога позволяет проводить быстрый поиск данных по различным критериям, и получать атрибутивную информацию об отдельном объекте, включая ссылку на данные. Использование технологии каталогизации метаданных на основе международных стандартов позволяет интегрировать внешние источники данных, использующие данный подход предоставления данных. Для описания метаданных геологических карт используется профиль метаданных по стандартам ISO 19115 и ISO 19139. Данные, представленные в каталоге, хранятся у поставщика данных на внешнем узле в виде векторных файлов, а также в виде отдельных слоёв в рамках сервисов доступа к пространственным данным, таких как OGC Web Map Service (OGC WMS) и OGC Web Feature Service (OGC WFS). Для реализации сервиса каталога используется программный пакет с открытым исходным кодом GeoNetwork. На данный момент в каталоге доступны метаданные Всероссийского Геологического Института (ВСЕГЕИ) масштаба 1:1000000 третьего поколения по территории России (131 запись), а также метаданные ВСЕГЕИ масштаба 1:200000 второго поколения по территории России (212 записей).

В Среде организован **удаленный доступ к метаинформации о месторождениях РФ и государственных геологических отчетах, находящихся в БД «Роснедра»**, которые содержат информацию о 52 тыс. месторождений и 478 тыс. геологических отчетах.

Поиск по месторождениям использует фасетную технологию. Пользователь может выбрать месторождение по наименованию (после ввода первых четырех символов всплывает подсказка), или выбрать область, населенный пункт и т.п. Отдельно выбирается тип полезных ископаемых (на основе всплывающей подсказки). Данное решение является компромиссом между вводом полных наименований и выводом многостраничного списка. Таким образом, поиск осуществляется выбором не по индексу, а полнотекстовым перебором, что позволяет выполнять его с использованием регулярных выражений.

Спутниковый блок предоставляет пользователям единую точку входа к данным спутников Aqua, Terra, Landsat, orbview-3 и к другим мультиспектральным данным высокого и среднего разрешения. Источником данных служат порталы спутниковых данных Центра спутникового мониторинга Института автоматики и процессов управления ДВО РАН, NASA, Геологической службы США (USGS). Поиск данных осуществляется в одном из трех режимов: поиск с помощью пользовательского поиска внешнего портала, поиск по метаданным, поиск по собственной базе данных метаданных спутниковых снимков.

Спутниковые снимки обрабатываются в зависимости от имеющейся информации, но в любом случае пользователям предоставляется вся информация, включая обзорные изображения.

Система поиска преобразует сформированный пользователем запрос в последовательность запросов поиска названий и поиска по географическим координатам. После этого, запросы последовательно, из браузера пользователя передаются поисковым машинам отдельных блоков Портала (машины обрабатывают запросы параллельно), а они, в свою очередь, либо осуществляют поиск самостоятельно, либо обращаются к собственным поисковым системам блоков Портала или к глобальным поисковым системам.

Интеграция разнородных БД метаданных из разных блоков Среды на единой платформе. Перспективной представляется интеграция описанного выше подхода с технологиями нового типа информационных систем, т.е. операций как с данными, так и с наборами данных. Объектом хранения таких систем являются наборы данных, т.е. таблицы. Новые данные в систему попадают через интерфейс пользователя или по протоколам обмена метаданными и данными OAI.

Вычислительная-аналитическая блок Среды - облачный инструментарий пользователей для обработки различных типов геологических данных. Предложенный при построении подход предполагает использование внешних вычислительных узлов для обработки данных, взаимодействие с которыми осуществляется с использованием технологии web-сервисов, в частности OGC Web Processing Service.

Реализованная платформа выступает в роли посредника между пользователем и внешними системами обработки, предоставляя единый интерфейс доступа ко всем алгоритмам обработки, имеющихся во внешних системах обработки (узлах системы). Описанная архитектура также предполагает возможность использования данных не только из доступных в системе открытых источников, но и загрузку данных для обработки самим пользователем. В разрабатываемой Среде вычислительно – аналитические блоки обработки и анализа геологической информации организованы в виде наборов служебных и аналитических функций с возможностью пользовательского доступа к выбору метода обработки; цепочек обработки, включающих загрузку данных, трансформацию форматов, методов анализа и визуализации

результатов; тематических цепочек, осуществляющих последовательность методов анализа. Доступ к сервисам обработки и анализа осуществляется через платформу управления распределенными сервисами обработки данных.

В настоящее время Вычислительно-аналитическая геологическая среда [9] включает в себя следующие узлы обработки:

- *Многомерные методы анализа данных.* Включает в себя набор методов для многомерного анализа количественных данных, таких как факторный анализ, кластерный анализ, регрессионный анализ и т.д. В качестве компонента для реализации модуля статистического анализа количественных данных был выбран язык программирования R. Интерфейс взаимодействия с сервисами построен с использованием модуля Rserve. Узел разработан и поддерживается в Государственном геологическом музее им. В.И.Вернадского РАН [10].
- *Обработка спутниковых данных.* Включает в себя методы первичной обработки спутниковых данных, такие как калибровка и пространственная привязка спутниковых данных. Узел разработан и поддерживается в Институте автоматики и процессов управления ДВО РАН.
- *Обработка петролого-геохимических данных.* В Институте Физики Земли РАН разработана интерактивная база методов обработки петролого-геохимических данных [2]. Система предоставляет сервисы построения спайдерграмм, гистограмм и классификационных диаграмм; сервис идентификации минералов по их химическому составу; сервис интерпретации состава минерала и разложение на миналы 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 [7]. Метаданные включают в себя авторов, принадлежность организации, абстракт, 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 [8]. 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.

Для обеспечения высокого уровня надёжности работы сервисов в рамках вычислительно-аналитической среды разработана **система мониторинга**, позволяющая оперативно реагировать на изменения в работе сервисов [1]. Использование разнородных сервисов, взаимодействие с которыми осуществляется с помощью различных протоколов и по различным интерфейсам, подразумевает ряд ограничений на предмет мониторинга. Однако, имея общую техническую информацию о каждом сервисе (web-адрес сервиса, протокол, версия протокола и т.д.), можно реализовать следующие общие виды проверок:

- Проверка доступности удалённого узла;

- Проверка работоспособности сервиса на удалённом узле по требуемому протоколу взаимодействия;
- Проверка наличия изменений в работе сервиса на основе тестовых запросов к WPS процессам.

Более сложные виды проверок состояния сервисов требуют знания детального описания интерфейса взаимодействия, что делает такой вид проверок зависимым от конкретной реализации сервиса.

Используя описанные методы проверки состояния сервисов, можно формировать статистику доступности отдельных сервисов. В случае проблем с доступом к определённому сервису, можно предлагать пользователям альтернативные реализации подобного сервиса, в случае их наличия в среде.

Перспективы развития. Объединение тематических ресурсов в общую интегрированную информационную инфраструктуру поддержки научных исследований позволит получить прямой доступ ко всем узлам Системы, уменьшит количество действий пользователя при общении с каждым узлом, позволит отдельным узлам, так и отдельным сервисам данных узлов быть интегрированными в другие информационные системы, облегчит процесс администрирования Системы.

Для решения этой и других задач, таких как: создание общей БД метаданных в едином программном формате, получение статистики о ресурсах Среды, разработка достаточно простого общего поискового механизма ко всем информационным блокам и возможности передачи результатов поиска в настоящее время авторами разрабатывается подход, позволяющий интегрировать разнородные базы мета-данных из различных информационных блоков Среды в единую подсистему на платформе CKAN. CKAN - мощная система управления данными с открытым программным кодом, которая делает данные доступными, обеспечивая инструменты для оптимизации данных, их совместного использования, нахождения, представления и хранения. Платформа CKAN относится к новому типу информационных систем - систем управления данными (DMC), основанных на принципах "открытого доступа" и работах CODATA. Объектом хранения такой системы являются наборы данных, т.е. таблицы. Новые данные в систему попадают через интерфейс пользователя или по протоколам обмена метаданными и данными OAI.

ЛИТЕРАТУРА

[1] *Ерёменко В. С., Наумова В. В.* Система каталогизации и мониторинга территориально распределенных вычислительных узлов в среде WPS сервисов для решения геологических задач // Вестник НГУ. Серия: Информационные технологии. 2019. Т. 17, No 2. С. 39–48. DOI 10.25205/1818-7900-2019-17-2-39-48.

[2] *Иванов С.Д.* Интерактивный реестр геосенсоров на основе веб-приложения // Компьютерные исследования и моделирование, 2016. Т. 8. № 4. С. 621–632.

[3] *Наумова В.В., Горячев И.Н, Дьяков С.Е., Белоусов А.В., Платонов К.А.* Современные технологии формирования информационной инфраструктуры для поддержки и сопровождения научных геологических исследований на Дальнем Востоке России // Информационные технологии, 2015 г., №7 - С. 551-559

[4] *Наумова В.В.* Информационные и функциональные возможности информационной интернет-инфраструктуры для поддержки научных исследований в области геологии //XVI Российская конференция «Распределенные информационно - вычислительные ресурсы. Наука – цифровой экономике» (DICR-2017): Труды XVI Всероссийской конференции (4-7 декабря 2017 г.). Новосибирск / Под ред. О.Л. Жижимова, А.М. Федотова. - 2017. - Новосибирск: ИВТ СО РАН. - С.44-49

- ISBN: 978-5-905569-10-4.

[5] *Шокин Ю.И., Федотов А.М.* К вопросу о развитии информационной инфраструктуры СО РАН // Вычислительные технологии. - 2009. - Т.14. - № 6. - С.127-137

[6] *Шокин Ю. И., Федотов А. М., Жижимов О. Л.* Технологии создания распределенных информационных систем для поддержки научных исследований // Вычислительные технологии. - 2015. -Т. 20, № 5. - С. 251-274

[7] *Dominika Tkaczyk, Pawel Szostek, Mateusz Fedoryszak, Piotr Jan Dendek and Lukasz Bolikowski*. CERMINE: automatic extraction of structured metadata from scientific literature. In International Journal on Document Analysis and Recognition, 2015, vol. 18, no. 4, pp. 317-335, doi: 10.1007/s10032-015-0249-8.

[8] *Diana Maynard, Kalina Bontcheva, and Isabelle Augenstein.* Synthesis Lectures on the Semantic Web: Theory and Technology, December 2016, Vol. 6, No. 2 , Pages 1-194

[9] *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, vol. 18, no. 6 (current), 2018. DOI: 10.2205/2018ES000636

[10] *Platonov K.A.* Methods and Technologies for Integration and Processing of Territorially Distributed Quantitative Geological Information // Proceedings of the XX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2018), Moscow, Russia, October 9-12, 2018, p. 348-353.