

УДК 004.550

ЦИФРОВОЙ РЕПОЗИТОРИЙ "GEOLOGYSCIENCE.RU": ОТКРЫТЫЙ ДОСТУП К НАУЧНЫМ ПУБЛИКАЦИЯМ ПО ГЕОЛОГИИ РОССИИ

М. И. Патук¹, В. В. Наумова², В. С. Ерёменко³

Федеральное государственное бюджетное учреждение науки «Государственный геологический музей им. В.И. Вернадского РАН»

¹patuk@mail.ru, ²naumova_new@mail.ru, ³vitaer@gmail.com

Аннотация

Описаны новые подходы, связанные со сбором данных из разнородных информационных систем доступа к научным публикациям с использованием открытых международных стандартов и протоколов для формирования систем открытого доступа к научным геологическим публикациям. На основе разработанных и адаптированных подходов и технологических решений реализован комплекс программ информационно-аналитической системы доступа к научным публикациям, реализующей функции сбора, поиска, каталогизации, фильтрации и управления научными публикациями и их метаданными.

Ключевые слова: *информационные технологии, науки о Земле, репозиторий, научные публикации.*

Введение

Термин «открытый доступ» впервые был упомянут на Будапештской конференции по открытому доступу в феврале 2002 г. [1]. С тех пор его смысл практически не изменился: Open Access определяется как бесплатный (free), оперативный (immediate), постоянный (permanent), полнотекстовый (fulltext), онлайн-доступ (online) к научным публикациям.

Открытый доступ — это бесплатный доступ читателей к онлайн-научным публикациям с правом читать, загружать, копировать, распространять, печатать, искать, ссылаться на полнотекстовые статьи, индексировать и т. п., то есть использовать с любой законной целью без финансовых, юридических или технических препятствий.

Выделяют два основных технологических направления: журналы открытого доступа и архивы (репозитории) открытого доступа. Оба направления – способы научного общения. Журналы открытого доступа публикуют прореферированные статьи, а репозитории собирают документы – не обязательно прошедшие реферирование и не обязательно статьи. Репозиторий собирает «свои» работы, т. е. труды сотрудников данного учреждения, и этим принципиально отличается от библиотеки. Журналы открытого доступа и репозитории не являются взаимоисключающими – они дополняют друг друга.

Цифровые репозитории служат реальным показателем качества деятельности университета/института и показывают научную, социальную и экономическую значимость исследовательских работ и таким образом демонстрируют статус и общественное значение университета/института.

Для практической реализации моделей открытых архивов предполагается:

- обмениваться метаданными, а не самими цифровыми объектами;
- использовать асинхронную технологию сбора данных;
- сформировать две группы участников системы «Открытого архива»: поставщики данных и поставщики услуг.

Поставщики данных (открытые архивы, репозитории) обеспечивают свободный доступ к метаданным и бесплатный доступ и пользование ресурсами, а также простоту в работе, не требующую создания каких-либо специальных коллективов и поэтому открывающую двери для участия малых организаций,

Поставщики услуг используют интерфейсы открытых архивов поставщиков данных, собирают и хранят метаданные, выбирают некоторые специализированные коллекции от поставщиков данных, пополняют состав метаданных и обогащают метазаписи, обеспечивают обслуживание на основе метаданных.

Разработано и успешно работает бесплатное программное обеспечение с открытым кодом для создания и поддержки таких OAI-совместимых архивов: наиболее популярные E-prints (<https://www.eprints.org/>), Dspace (<http://dspace.org/>).

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»: консор-

лидация 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 [2].

В последнее десятилетие наблюдается качественно новый уровень в организации хранения и предоставления «открытых научных данных». Активно развиваются системы и платформы, которые обеспечивают весь процесс управления данными – от публикации автором до анализа и повторного использования этих данных любым исследователем или системой. В концептуальной основе новых систем лежат «принципы цитирования данных» DataCite [3], FAIR-принципы [4] и рекомендации мировых ассоциаций по обмену данными: The International Science Council (ISC) [5], The Research Data Alliance (RDA).

Современной формой хранения и предоставления научной информации стали наборы данных. Набор данных – это контейнер, содержащий данные, метаописание (в формате Dublin Core или DataCite) и уникальный идентификатор (например, DOI). Мировые центры данных предоставляют доступ к своим хранилищам по протоколам Open Archives Initiative (OAI). Организация научной информации в виде наборов данных и доступность их метаданных по протоколам OAI позволяет упростить автоматизацию процессов поиска геологической информации по России. Надежность данных подтверждается указанием авторства, выходных данных статьи, проектов, программ, в рамках которых проводились исследования.

В рамках разработки Информационно-аналитической геологической среды "GeologyScience.ru", которая обеспечивает единую точку доступа к геологическим данным и системам их обработки на территории России [6–8], создан и поддерживается блок управления научными публикациями – <http://repository.geologyscience.ru/> (Рис. 1).

Блок разрабатывается как самостоятельный проект, но обладает свойствами для интеграции его в территориально-распределенные системы: однородность данных, наличие базы метаданных в международных форматах, доступ через API, поддержка сквозной авторизации и разграничения прав, службы

мониторинга и статистики. На Рис. 2 показано, каким образом блок «Научные публикации» входит в Информационно-аналитическую геологическую Среду.

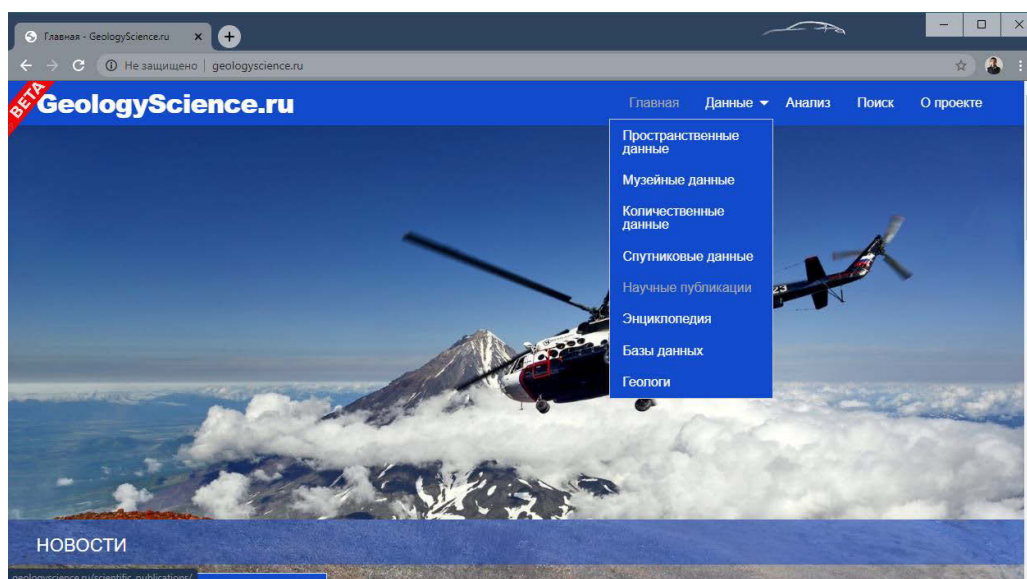


Рисунок 1. Главная страница портала “GeologyScience.ru”.

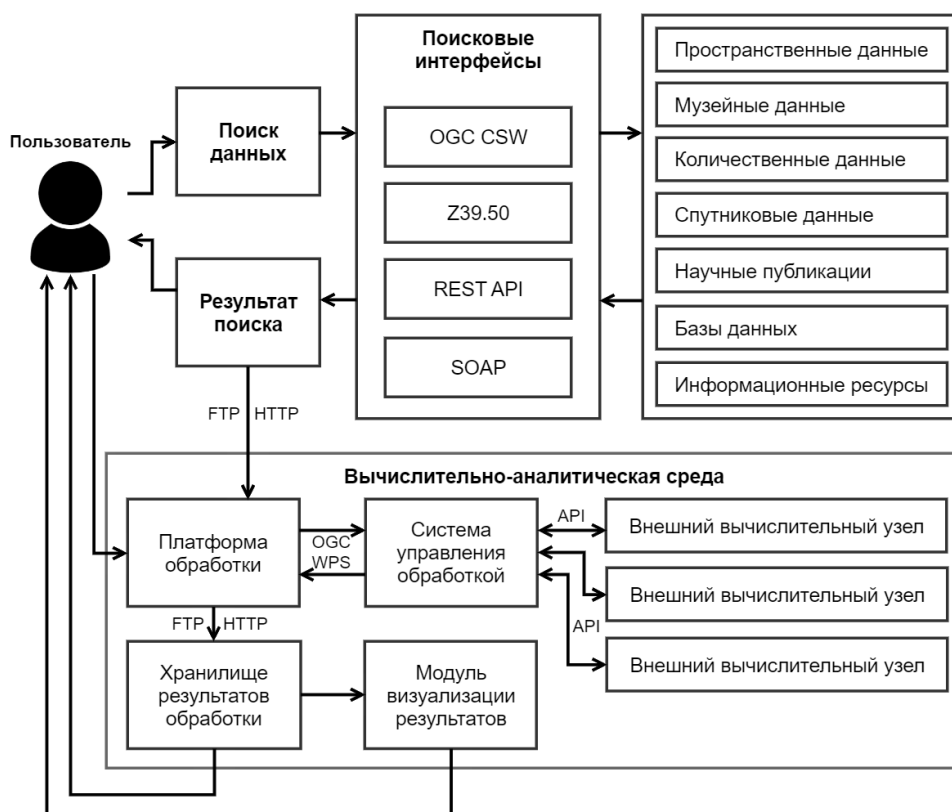


Рисунок 2. Обобщенная функциональная схема Информационно-аналитической геологической среды.

Авторами кратко сформулированы задачи блока «Научные публикации»:

- сбор и интеграция научных публикаций из территориально-распределенных источников;
- хранение данных метаданных в машиночитаемом виде и в принятых международных форматах;
- обеспечение доступности данных для пользователей и программ;
- тематическая адаптация интерфейса и функций системы.

Репозиторий «Геология России» создан на основе свободно распространяемого программного обеспечения DSpace 6.3. Системы подобного класса организованы для хранения и обмена цифровыми объектами и предоставляют доступ к информации по протоколу OAI. Имеются удобная система каталогов, наличие сервера сбора метаданных OAI-PMH, полнотекстовый поиск, основанный на поисковом инструменте Apache Lucene или Apache Solr, разграничение прав и поддержка протокола доступа LDAP, открытый код и большое сообщество пользователей и разработчиков по всему миру, возможность управлять и хранить цифровой материал любого типа.

Согласно общепринятому делению данный Репозиторий является тематическим. Тематика Репозитория – науки о Земле (геология, геохимия, петрология, минералогия, тектоника, геоморфология, вулканология, палеонтология, стратиграфия и т. п.). Основой информации для Репозитория являются научные статьи, монографии, диссертации, авторефераты диссертаций, тезисы докладов, материалы конференций, находящиеся в открытом доступе.

Коллекции из архивов и библиотечных каталогов могут быть доступны на интернет-ресурсе через три коммуникационных протокола: OAI-PMH (the Open Archives Initiative Protocol for Metadata Harvesting), Z39.50 или SRU (Search/Retrieve via URL).

Выбор коммуникационного протокола имеет большое влияние на функциональность, которую ресурс может предоставить конечному пользователю. Хотя все три протокола предоставляют стандарт для коммуникации между Репозиторием и библиотечными системами, коммуникационная парадигма, лежащая в основе, значительно различается. В то время как OAI-PMH позволяет ему собирать все записи метаданных из библиотек в центральный архив, Z39.50 и SRU

были разработаны для удаленного доступа и предоставления, поэтому записи метаданных остаются только у провайдера данных.

Сбор метаданных по протоколам Z39.50/SRU представляет значительную сложность. Протоколы Z39.50/SRU не предназначались для сбора метаданных, поэтому некоторая функциональность, требуемая для обеспечения эффективности и надежности процесса сбора, не была включена в проект протоколов.

В отличие от OAI-PMH, Z39.50-сервера доступны для значительного количества систем управления библиотеками, и его использование среди библиотек широко распространено. Большинство библиотек России в качестве системы управления библиотекой используется Ирбис (<http://irbis.elnit.org/>). Веб каталог системы, Web Ирбис, поддерживает сервер Z39.50, что позволяет попробовать осуществить сбор библиографических метаданных. В случае, когда сервер Z39.50 не настроен, можно воспользоваться функцией экспорта Web Ирбис, если она доступна.

Для поиска и извлечения информации из других репозиториев был создан скрипт на языке PHP. Обращение к сторонним репозиториям происходит по протоколу OAI_PMH.

Большинство репозиториев является институциональными, т. е. содержит информацию по многим направлениям, развиваемым в конкретном научном или образовательном учреждении, поддерживающем репозиторий.

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 текстовый файл с последующей загрузкой SQL-скриптом в таблицу PostgreSQL DSpace.

Данная информация позволяет в автоматическом режиме строить и обновлять предметный каталог репозитория, исходя из УДК русскоязычных публикаций по геологии России.

В текущий момент источниками данных для репозитория «Геология России» являются:

1. Репозитории:

- Репозиторий Томского политехнического университета. (<http://earchive.tpu.ru/>). Он является институциональным. Разделение по тематикам представлено слабо. Для отбора метаданных необходимо использовать словарь терминов. Платформа – DSpace.
- Репозиторий Института вулканологии и сейсмологии ДВО РАН. (<http://repo.kscnet.ru/>). Репозиторий тематический. Основная тематика связана с науками о Земле. В отборе метаданных по словарю терминов нет необходимости. Достаточно визуального контроля лог-файла импорта. Платформа – Eprints.
- Репозиторий Санкт-Петербургского государственного университета. (<https://dspace.spbu.ru>). Репозиторий институциональный. Четкое деление по тематикам, но незначительное количество данных. Отсюда взят раздел «Минералогия и кристаллография». Платформа – DSpace.

2. Электронные (цифровые) библиотеки:

- КиберЛенинка (<https://cyberleninka.ru/>). Открытая электронная библиотека. Четкое деление по тематикам. Присутствует раздел – Науки о Земле (earth-and-related-environmental-sciences). Требуется визуальный контроль лог-файла импорта из-за не всегда корректного отнесения публикаций к данному разделу. Данные можно брать из репозиториев по протоколу OAI-PMH. Но в отличие от репозиториев отсутствует сквозная нумерация данных, поэтому приходится сканировать данные по дате поступления.
- ELibrary (<https://www.elibrary.ru>). Самая представительная научная электронная библиотека. Но и самая требовательная к режиму доступа. В системе есть платное API для экспорта данных. Для реализации бесплатного доступа можно использовать свободно распространяемый скрипт – Subzer (<https://github.com/p1m-ortho/xs-sebzer>). Этот скрипт позволяет выгружать результаты поисковых запросов в формате BibTex. Система расширенного поиска в ELibrary достаточно функциональна, но, к сожалению, тематическим поиском пользоваться нельзя из-за

значительного количества некорректных результатов. Приходится пользоваться поиском по списку журналов и дате публикации.

3. PANGAEA. Сайт – <https://www.pangaea.de> – позиционирует себя как библиотека открытого доступа, предназначенная для хранения, публикации и распространения географически привязанных данных в науках о Земле. В основном представлены дополнительные данные (рисунки, фотографии, таблицы) к существующим публикациям. Как правило, имеется идентификатор DOI. В связи с тем, что представлены подробный предметный классификатор и ссылка на оригинал публикации, возможно использование этих данных для получения метаданных. Доступ к данным происходит по протоколу OAI-PMH. Данные сканируются по дате поступления.

МОДУЛИ ИМПОРТА ДАННЫХ

Для каждого из перечисленных источников создан свой модуль экспорта данных на языке PHP. Общая идеология работы схожа для всех модулей.

1. Сканирование источника:
 - сканирование по уникальным номерам (для репозиториев);
 - сканирование по дате поступления;
 - сканирование по списку предварительно отобранных ссылок.
2. Отбор данных:
 - выбор всех (если тематика – науки о Земле);
 - выбор по тематике (например – «earth-and-related-environmental-sciences»);
 - выбор по словарю терминов. Словарь терминов русскоязычный и содержит ~2100 терминов. Он создавался на основе ~2000 предварительно отобранных записей по интересующей нас тематике. Практика показала, что оптимальным оказалось 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.
 - для предотвращения дублирования импортируемых метаданных производится поиск наименования публикаций в базе данных DSpace – PostgreSQL. Дублированные метаданные исключаются из импорта.

3. Формирование XML-файлов со структурой DSpace Simple Archive Format с метаданными публикаций и экспорт файла публикации, при его наличии.
4. Импорт полученных данных в Репозиторий осуществляется стандартными средствами DSpace – импорт через простой архивный формат DSpace.

Анализ и обработка текстовых данных осуществляется в рамках Информационно-аналитической геологической среды "GeologyScience.ru", которая обеспечивает не только единую точку доступа к геологическим данным, но и к системам их обработки через Вычислительно-аналитическую геологическую среду <http://service.geologyscience.ru/> (Рис. 3) [9, 10].

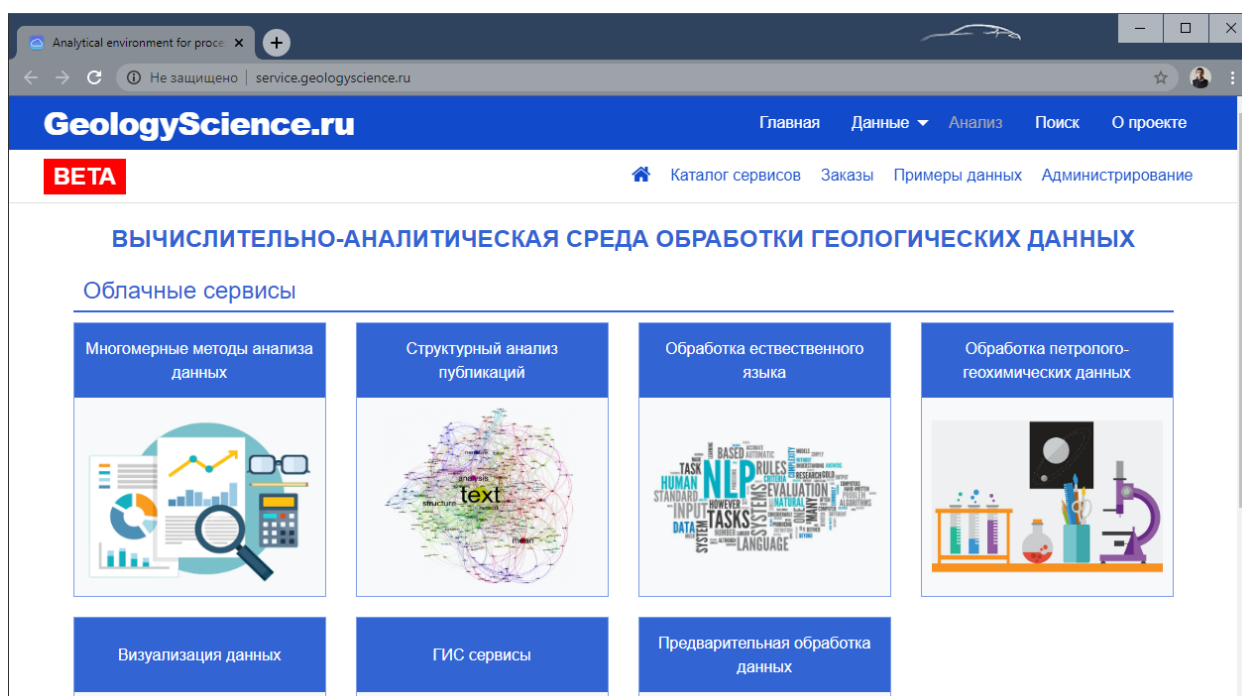


Рисунок 3. Единая точка входа к вычислительно-аналитической среде, входящей в состав "GeologyScience.ru".

В рассматриваемой вычислительно-аналитической среде для обеспечения единого подхода к вызову территориально-распределенных сервисов обработки и анализа данных используется стандарт OGC Web Processing Service (OGC WPS) в качестве промежуточного интерфейса.

Для анализа текстов в вычислительной среде доступны два удаленных вычислительных узла.

– **Структурный анализ публикаций.** В междисциплинарном центре математического и вычислительного моделирования (Университет Варшавы, Польша) разработан сервис для извлечения метаданных из научных публикаций [11]. Метаданные включают в себя авторов, аффилицию, 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 [12]. 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.

ЗАКЛЮЧЕНИЕ

Разработаны и адаптированы новые подходы, связанные со сбором разнородных данных из разнородных информационных систем доступа к научным публикациям с использованием открытых международных стандартов и протоколов в условиях системы ограничений, подразумевающей тематические, территориальные, временные ограничения, и методы, основанные на открытых международных стандартах метаданных научных публикаций и протоколов их обмена, теории построения распределенных информационных систем и поиска информации, для формирования тематических систем открытого доступа к научным публикациям.

Предложены новые технологические решения: сбора метаданных из разнотипных информационных систем, фильтрации собираемой информации с использованием УДК или специализированных тезаурусов, географического поиска на карте, организация тематического каталога.

На основе разработанных и адаптированных подходов и технологических решений реализован комплекс программ информационно-аналитической си-

стемы доступа к научным публикациям, реализующей функции сбора, поиска, категоризации, каталогизации, фильтрации и управления научными публикациями и их метаданными.

Благодарности

Научные исследования выполняются в рамках Государственного задания ФГБУН Государственного геологического музея им. В.И. Вернадского РАН по теме № 0140-2019-0005 «Разработка информационной среды интеграции данных естественнонаучных музеев и сервисов их обработки для наук о Земле».

СПИСОК ЛИТЕРАТУРЫ

1. Будапештская инициатива «Открытый доступ» / Budapest Open Access Initiative [Электронный ресурс]
URL: <https://www.budapestopenaccessinitiative.org> (дата обращения: 23.05.2020).
2. Технология открытых архивов / Open Archive Initiative, OAI [Электронный ресурс] URL: <https://openarchives.org/> (дата обращения: 23.05.2020).
3. *Brase J.* DataCite – A global registration agency for research data // In: Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology. IEEE, Beijing. 2009. P. 257–261. DOI: 10.1109/COINFO.2009.66
4. *Wilkinson M.D. et al.* The FAIR Guiding Principles for scientific data management and stewardship // In: *Sci. Data* 3:160018. 2016. P. 1. DOI: 10.1038/sdata.2016.18.
5. *Claudia Emerson, Elaine M. Faustman, Mustapha Mokrane, Sandy Harrison.* World Data System (WDS) Data Sharing Principles // Zenodo. 2015. <http://doi.org/10.5281/zenodo.34354>
6. *Naumova V.V., Belousov A.V.* Digital repository «Geology of the Russian Far East» – an open access to the spatially distributed online scientific publications // *Russian Journal of Earth Sciences*. 2014. Vol. 14, No. 1. P. 1–8.
7. *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 // Труды XVII Международной конференции «Распределенные информационно-вычислительные ресурсы. Цифровые двойники и большие данные (DICR-2019)», 2019. С. 139–147.

8. *Naumova V.V., Eremenko V.S., Platonov K.A., Dyakov S.V., Patuk M.I., Eremenko A.S.* Development of geographically distributed information-analytical geological environment // *Russian Journal of Earth Sciences*. 2019. Vol. 19, No. 6. DOI:10.2205/2019ES000696.

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*. 2018. Vol. 18, no. 6 (current). DOI: 10.2205/2018ES000636

10. *Eremenko V.S., Naumova V.V.* Computational and Analytical Environment for Processing and Analysis of Geological Data // *Proceedings of the V International Conference "Information Technologies in Earth Sciences and Applications for Geology, Mining and Economy (ITES&MP-2019)"*, Moscow, Russia, October 14–18, 2019. Published on CEUR Workshop Proceedings (CEUR-WS.org, ISSN 1613-0073). Vol-2527. Edited by Vera V. Naumova, Aleksandr S. Eremenko. P. 14–19.

11. *Tkaczyk D., Szostek P., Fedoryszak M., Dendek P., Bolikowski L.* CERMINE: automatic extraction of structured metadata from scientific literature. In *International Journal on Document Analysis and Recognition*. 2015. Vol. 18, no. 4. P. 317–335. DOI: 10.1007/s10032-015-0249-8.

12. *Maynard D., Bontcheva K., Augenstein I.* *Synthesis Lectures on the Semantic Web: Theory and Technology*, December 2016. Vol. 6, No. 2. P. 1–194.

DIGITAL REPOSITORY "GEOLOGYSOURCE.RU": OPEN ACCESS TO SCIENTIFIC PUBLICATIONS ON RUSSIAN GEOLOGY

Mikhail I. Patuk¹, Vera V. Naumova², Vitaliy S. Eremenko³

Vernadsky State Geological Museum RAS, Moscow (Russia)

¹patuk@mail.ru, ²naumova_new@mail.ru, ³vitaer@gmail.com

Abstract

The article describes new approaches related to the collection of data from heterogeneous information systems of access to scientific publications using open international standards and protocols for the formation of systems of open access to scientific geological publications. Based on developed and adapted approaches and technological solutions, a set of programs of information and analytical system of access to scientific publications has been implemented, implementing functions of collection, search, cataloguing, filtering and management of scientific publications and their metadata.

Keywords: *Information technology, Earth sciences, repository, scientific publications.*

REFERENCES

1. Budapest Open Access Initiative [electronic resource] URL: <https://www.budapestopenaccessinitiative.org> (date of the application: 23.05.2020).
2. Open Archive Initiative, OAI [electronic resource] URL: <https://openarchives.org/> (date of the application: 23.05.2020).
3. *Brase J.* DataCite – A global registration agency for research data // In: Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology. IEEE, Beijing. 2009. P. 257–261. DOI: 10.1109/COINFO.2009.66.
4. *Wilkinson M.D. et al.* The FAIR Guiding Principles for scientific data management and stewardship // In: *Sci. Data* 3:160018. 2016. P. 1. DOI: 10.1038/sdata.2016.18.
5. *Claudia Emerson, Elaine M. Faustman, Mustapha Mokrane, Sandy Harrison.* World Data System (WDS) Data Sharing Principles // Zenodo. 2015.

<http://doi.org/10.5281/zenodo.34354>.

6. Naumova V.V., Belousov A.V. Digital repository «Geology of the Russian Far East» – an open access to the spatially distributed online scientific publications // Russian Journal of Earth Sciences. 2014. Vol. 14, № 1. P. 1–8.

7. Naumova V.V., Platonov K.A., Eremenko V.S., Patuk M.I., Dyakov S.V. Informacionno-analiticheskaja sreda dlja podderzhki nauchnyh issledovanij v geologii: tekushhee sostojanie i perspektivy razvitija // Trudy XVII Mezhdunarodnoj konferencii «Raspredelennye informacionno-vychislitel'nye resursy. Cifrovye dvojniki i bol'shie dannye (DICR-2019)», 2019. S. 139–147.

8. Naumova V.V., Eremenko V.S., Platonov K.A., Dyakov S.V., Patuk M.I., Eremenko A.S. Development of geographically distributed information-analytical geological environment // Russian Journal of Earth Sciences. 2019. Vol. 19, No. 6. DOI:10.2205/2019ES000696.

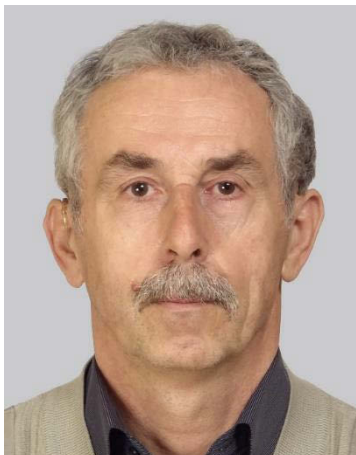
9. Eremenko V.S., Naumova V.V., Platonov K.A., Dyakov S.E., Eremenko A.S. The main components of a distributed computational and analytical environment for the scientific study of geological systems // Russian Journal of Earth Sciences. 2018. Vol. 18, no. 6 (current). DOI: 10.2205/2018ES000636.

10. Eremenko V.S., Naumova V.V. Computational and Analytical Environment for Processing and Analysis of Geological Data // Proceedings of the V International Conference “Information Technologies in Earth Sciences and Applications for Geology, Mining and Economy (ITES&MP-2019)”, Moscow, Russia, October 14–18, 2019. Published on CEUR Workshop Proceedings (CEUR-WS.org, ISSN 1613-0073). Vol-2527. Edited by Vera V. Naumova, Aleksandr S. Eremenko. P. 14–19.

11. Tkaczyk D., Szostek P., Fedoryszak M., Dendek P., Bolikowski L. CERMINE: automatic extraction of structured metadata from scientific literature. In International Journal on Document Analysis and Recognition. 2015. Vol. 18, no. 4. P. 317–335. DOI: 10.1007/s10032-015-0249-8.

12. Maynard D., Bontcheva K., Augenstein I. Synthesis Lectures on the Semantic Web: Theory and Technology, December 2016. Vol. 6, No. 2. P. 1–194.

СВЕДЕНИЯ ОБ АВТОРАХ



ПАТУК Михаил Иванович – к. г.-м. н., и. о. н. с., Научный отдел Государственного геологического музея им. В.И. Вернадского РАН, Москва, Россия.

Michail I. PATUK – PhD, scientific department, Vernadsky State Geological Museum RAS, Moscow (Russia).

Email: patuk@mail.ru



НАУМОВА Вера Викторовна – д. г.-м. н., г. н. с., зав. Научным отделом Государственного геологического музея им. В.И. Вернадского РАН, Москва, Россия.

Vera V. NAUMOVA – Prof., head of SGM scientific department, Vernadsky State Geological Museum RAS, Moscow (Russia).

Email: Naumova_new@mail.ru



ЕРЁМЕНКО Виталий Сергеевич – и. о. м. н. с., Научный отдел Государственного геологического музея им. В.И. Вернадского РАН, Москва, Россия.

Vitaliy S. EREMENKO – scientific department, Vernadsky State Geological Museum RAS, Moscow (Russia).

Email: vitaer@gmail.com

Материал поступил в редакцию 28 мая 2020 года