The Center of quantitative data on geology: Current state and prospects for development

K. A. Platonov[1] and V. V. Naumova[1]

Received 08 June 2020; accepted 25 November 2020; published 7 December 2020.

"Open access" to scientific data led to develop new approaches in data management. The study describes methods and approaches for creation of subject geological system of quantitative information management. The types of existing open sources of quantitative data sets are considered. Examples of sources of data sets of various types with a brief description are given. The international standards used in the construction of the system are described. The proposed technological solution is based on principles of DataCite, FAIR-systems and OAI protocols. The system allows to automatically generate data collections from geographically distributed sources: repositories, databases, world data centers. The system is being developed as part of the work on creating an Information-Analytical Geological Environment, which provides a single access point to geological data on the territory of Russia and their processing systems. The software modules and solutions that were used to create the system based on the proposed approach are described. The subject adaptation of the interface is also described.
KEYWORDS: Geographically distributed sources; quantitative datasets; geology of Russia.

Citation: Platonov, K. A. and V. V. Naumova (2020), The Center of quantitative data on geology: Current state and prospects for development, *Russ. J. Earth. Sci., 20,* ES6012, doi:10.2205/2020ES000755.

Introduction

The last decade shows qualitatively new level of storage and provision of "open scientific data". The systems and platforms that provides entire process of data management from publication by the author to the analysis and reuse of this data by any researcher or system are actively developing. The new systems are conceptually based on the "principles of data citation" DataCite [*Brase,* 2009], FAIR-principles [*Wilkinson et al.,* 2016] and recommendations of world data exchange associations: The International Science Council (ISC) [*Emerson et al.,* 2015], The Research Data Alliance (RDA).

Datasets have become a modern form of scientific information storage and presentation. A data set is a container containing data, meta information (in a format of Dublin Core or DataCite) and unique identifier (e.g. DOI). World datacenters provide access to their storages using Open Archives Initiative (OAI) protocols. Organization of scientific information in a form of data sets and accessibility of its metadata using OAI protocols allows to simplify automation of searching processes of geological information about Russia. Getting access to these data is an important task for the geological exploration of the territory of Russia. The reliability of data is confirmed by the indication of authorship, the output data of the article, projects, programs in the framework of which the studies were conducted.

The Project "GeologyScience.ru" – the Block of Quantitative Data

Within the development of Information-analytical environment "GeologyScience.ru", that provides unified access point to geological

[1]V. I. Vernadsky State Geological Museum, Moscow, Russia

http://rjes.wdcb.ru/doi/2020ES000755-res.html

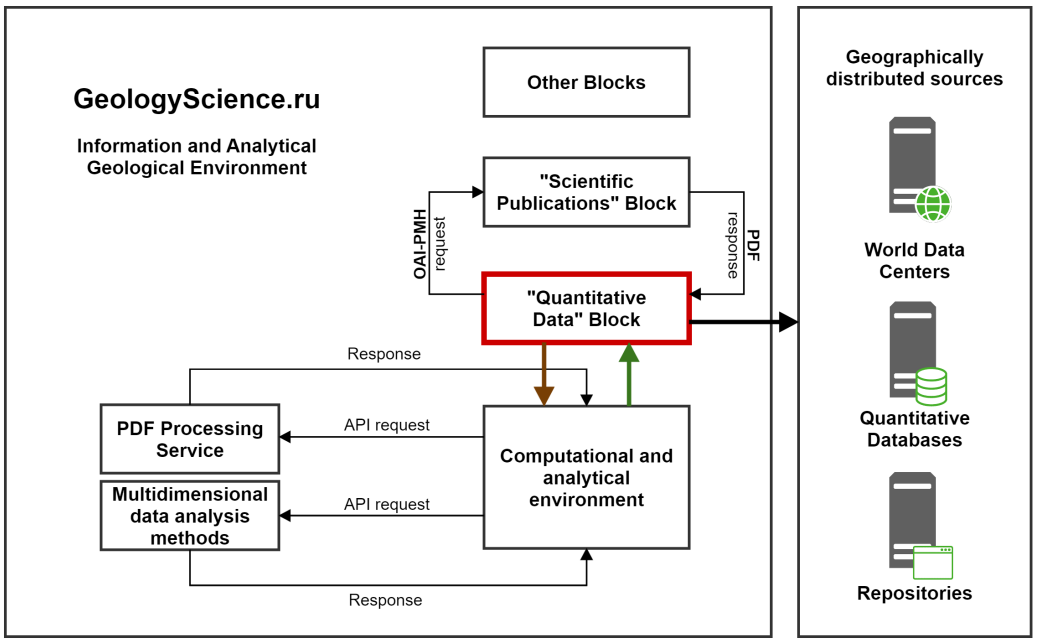

Figure 1. Conceptual scheme of "quantitative data" block layout in the Environment.

data of Russian territory and systems of its processing [*Naumova et al.*, 2019; *Eremenko et al.*, 2018] a block of quantitative information management is created and supported (http://datacenter.geologyscience.ru).

The Center is developing as standalone project but having features for its integration into geographically distributed systems: homogeneity of data, existence of international formats database, API access, end-to-end authorization support and access control, monitoring and statistic services [*Shokin et al.*, 2015]. Figure 1 shows how the "Quantitative Data" block is included in the Information-Analytical Geological Environment.

In the previous studies [*Platonov*, 2015; *Platonov and Naumova*, 2017] basic modules of the system and principles of forming of primary base of data sets on the basis of quantitative data tables extracted from scientific publications are described (Figure 2).

With the development of the availability of "open scientific data", the number of sources suitable for automatic processing for information about Russia is growing. The necessity of development of methods and technologies for integration, management and cataloging of quantitative data sets from these sources is growing.

The objectives of the "quantitative information" block are briefly formulated:

- Collection and integration of quantitative data from geographically distributed sources;
- Storage of machine-readable data and metadata in accepted international formats;
- Provision of accessibility of data for users and programs;
- Thematic adaptation of interface and functions of the system.

Sources of Quantitative Data Sets

It is possible to distinguish three main sources: repositories that store scientific publications texts, geological quantitative databases at institutes and scientific institutions, world data centers that act as publisher systems of full sets of scientific data.

Sources such as social networks and journal systems contain tables of quantitative information but are not suitable for automatic processing for several reasons. For instance, ResearchGate social network does not support long-term storage. The mechanisms of exchange and export are also absent. And

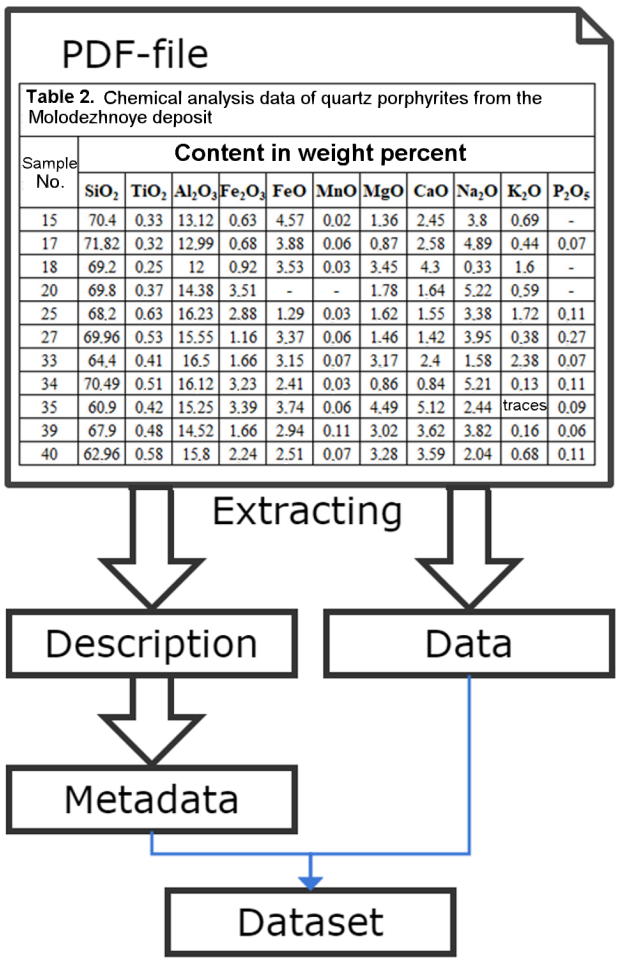

Figure 2. Principles of data sets formation from scientific publication.

in the Elsevier journal system, data sets are part of the publication and do not have their own description and the ability to search for them.

The following systems were selected as the first sources.

Digital repository "Geology of Russia" (http://repository.geologyscience.ru), which integrates scientific publications into the created "GeologyScience.ru" Environment [*Naumova and Belousov,* 2014; *Naumova et al.,* 2015]. The repository is based on free software DSpace and contains more than 2000 of full text scientific publications related to Earth science. Systems of this class are organized for the storage and exchange of digital objects and provide access to information via the OAI protocol.

The World Data Center "PANGAEA" [*Diepenbroek,* 2002], developed and maintained by the Alfred Wegener Institute for Polar and Marine Research. The portal contains over 368,000 data sets published as scientific papers. The PANGAEA project supports FAIR-principles and provides programmatic access to metadata and data using the OAI protocol and REST interface.

The EarthChem portal maintained by Columbia University (http://www.earthchem.org) provides a single point to the quantitative databases of Earth sciences. One of the GeoRoc databases (`http://georoc.mpch-mainz.gwdg.de/georoc/`) contains quantitative information from 18,000 publications [*Sarbas,* 2008]. Access to the database is performed through a search interface based on the REST architecture.

Data Sets Integration

The global metadata scheme developed within the DataCite project [*Data*, 2014; *DataCite*, 2017], allows organizing information integration contained in geographically distributed sources at a logical level. The DataCite schema is an extension of the Dublin Core scheme. The scheme contains 3 levels of description: mandatory, recommended and extra fields. During integration of data from sources tasks of filtration, bilingual support (Russian and English), extraction of concepts and generation of a meta description arise (in case of its missing).

The process of automatically adding new data is as follows: The System periodically initiates a request to the service of OAI source for receiving of all or just new entries. Metadata are investigated for data belonging to the territory of Russia: detection of geographical concepts or presence of coordinates in metadata. After the filtration procedure there are two scenarios possible. If input metadata describes a dataset they are serialized as is in internal repository and a direct download link to table is created. Otherwise a new metadata in a format of DataCite are generated using received description. The description is a bibliographic link or metadata to publication source of quantitative table.

Automation of data collection from sources that support only an interface based on the REST architecture requires the creation of an algorithm for transforming the source response into a data set. For example, an author's algorithm for accessing the geochemical database "GeoROC" through the REST interface of the EarthChem integration portal was developed.

Software Implementation and Services

The Core of the System is developed using PHP in a form of set of Drupal 7.x platform modules. The platform allows to develop systems of any complexity using proprietary and third-party solutions. The Core of the System is responsible for tables of the Center database creation and for the logic of the interface, administrative parts and service components. It includes data filtration, cataloguing, statistics collection and sources monitoring, interaction with external services and systems, searching module.

For collection of metadata the PHP OAI PMH library proposed at site of Initiative of Open Data is used. The library provides necessary functionality for polling a source using the OAI-PHM protocol.

The database of metadata is formed and maintained using a set of ISLANDORA project modules (https://islandora.ca). The project is being developed as a program environment for organizing a collaborative management and digital objects investigation. In the basis of a storage system the Fedora open repository platform (Flexible Extensible Digital Object Repository Architecture) is used. The whole information about supported metadata schemes, the metadata, data, structure and links are stored in the repository. The interface part of the Project developed for Drupal allows to control digital objects, additional services and general settings.

Search functions and metadata indexing are performed using the Solr search platform.

The project provides access through an interface based on the REST architecture. Through the API and the query language, Solr searches for data sets in the system and returns a response in JSON format. Functions for creating, modifying, and deleting objects are also available.

ISLANDORA OAI module allows to provide content as objects of a digital repository. The module supports popular metadata schemes (MODS, DC, DATACITE) and transformation from one scheme to another.

To create files in spreadsheet format, the PHPExcel library is used.

For interactive work with tabular information on the site, the JavaScript library "Tabulator" (http://tabulator.info/) is used. In addition to data visualization functions, the library allows you to edit data on the fly and export the result in CSV and XLS format.

The Quantitative Data Center uses several external services: a pdf file processing service (http://gext.geologyscience.ru) [*Platonov,* 2018], a quantitative table processing node (http://service.geologyscience.ru/service?nodeId=1), Australian Scientific Dictionaries (https://vocabs.ands.org.au) and Google's automatic translation service (https://translate.google.com).

Initially, the first two services were part of the system being created. But the modern concept of "open access" involves the distribution of not only information, but also the services of its processing. Therefore, in the structure of the Center, these parts were transformed into independent nodes and became part of the Computational-analytical geological environment. Now any user or system can apply and get the results of the nodes. The transition to the use of external processing nodes by the System makes it easy to add or change them to more suitable ones without changing the structure.

Interface of the System

The user interface is thematic, i.e. it uses concepts and services that it can understand, and to which a user geologist can quickly adapt. The provision of information in the system is divided into four categories: "What?" – search by name, "Where?" – spatial search (in development), "When?" – temporary search, "Who?" – search by personnel. In each category, full-text search and selection of values by catalog is available (Figure 3).

In the category of "What?" The catalog of mineral deposits "ROSGEOLFOND" and the catalog of geological objects of Russia (supplemented) are used. Category "Where?" In addition to searching on the map, it allows you to search by geographical names. Accordingly, the category "When?" understands the values from the stratigraphic scale. And the category of "Who?" contains the names of the authors of the data set.

To automatically bind sets to directory values, metadata is processed for the search for concepts.

Most metadata are described in the Dublin Core format. The values of the meta-fields "Creator", "Contributor" are used to form the list of authors. The "Coverage" field contains geographical names or coordinates. The "Subject" field contains a list of

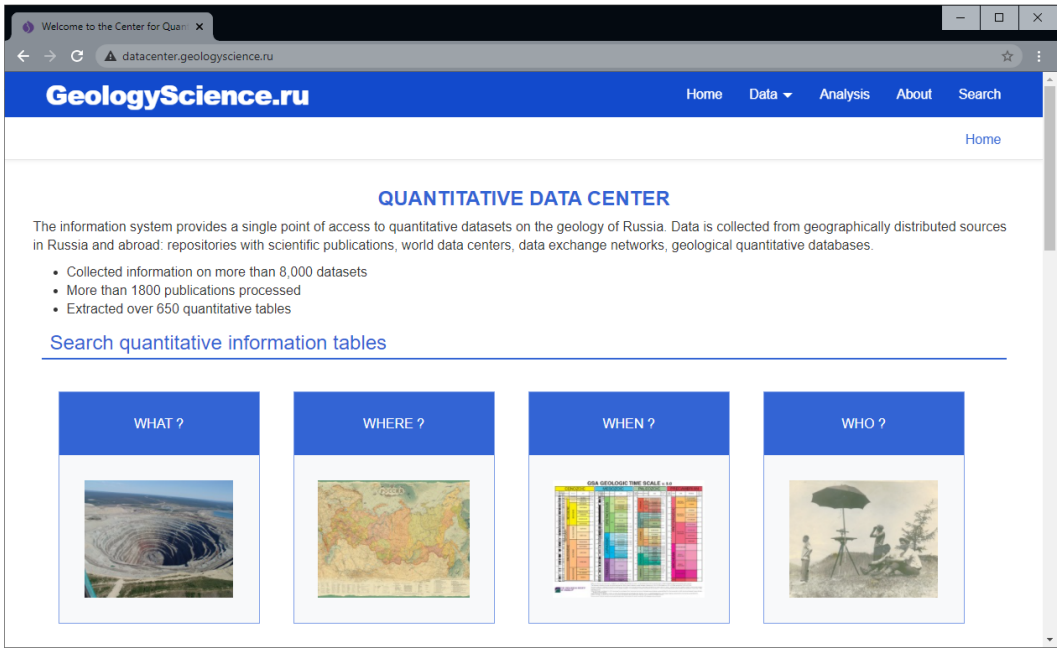

Figure 3. Main page of the Center of quantitative data.

keywords that may belong to any of the directories. Unstructured fields "Title", "Description", "Source", "Relation" are subjected to additional processing to detect concepts. Basically, this is a search for fixed values in the text of meta-fields. For metadata in the DataCite format, the "GeoLocation" and "FundingReference" fields are checked.

The additional complexity of automatic cataloging is in English. In the metadata in English, Russian names are recorded in transliteration. To support two languages, the Google Translate machine translation system and the "Russian-Latin/BGN/PCGN" transliteration system are used (Figure 4).

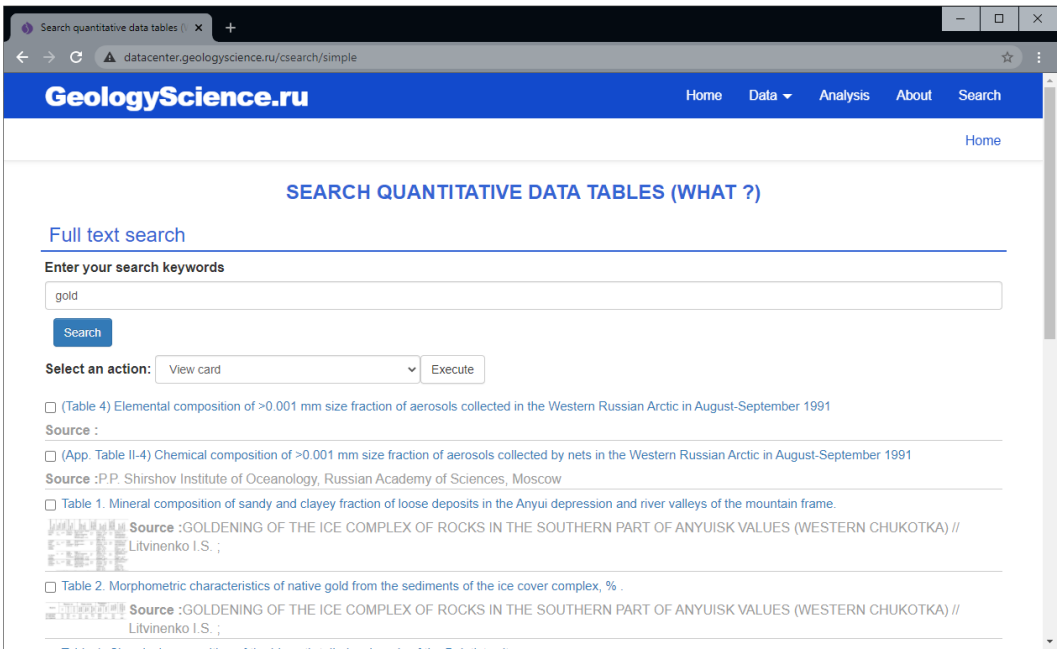

Figure 4. Example of the search answer with bilingual support.

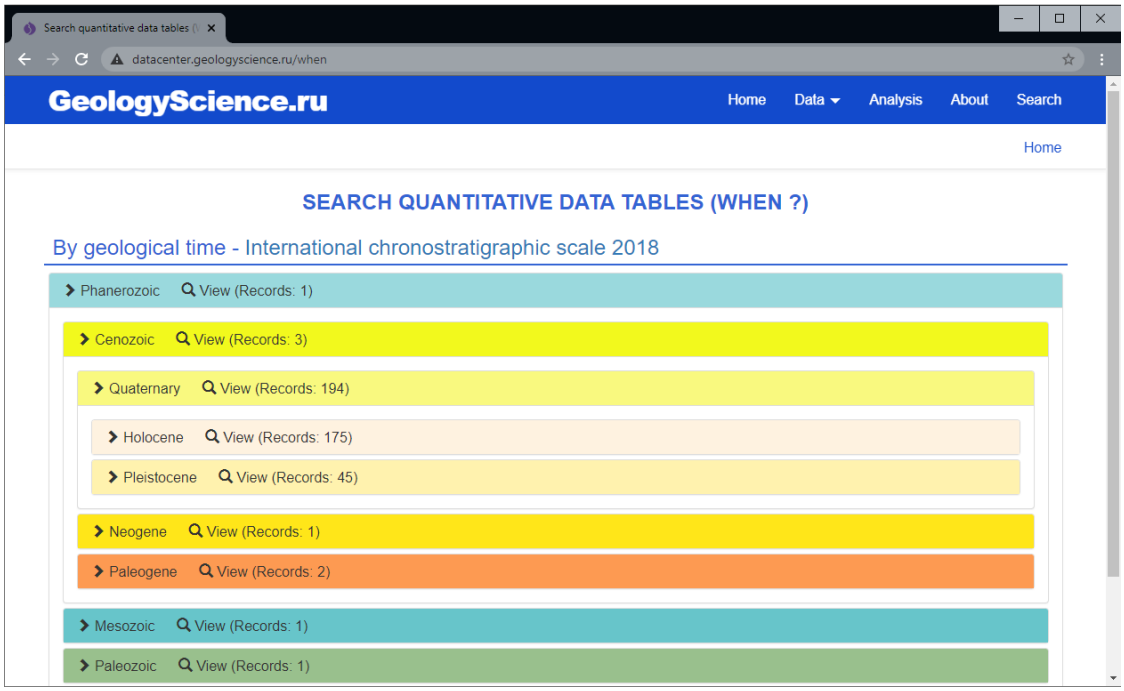

Figure 5. Datasets search on the catalogue of international stratigraphic scale.

To create a catalog of concepts of the stratigraphic scale and absolute age values, the Research Vocabularies Australia service is used. In the form of an RDF schema, the Geological Timescale 2018 dictionary is published by the International Commission on Stratigraphy. The dictionary contains values in 6 languages and allows you to bring old data to modern meanings. The result of the automatic linking of datasets to the stratigraphic scale is shown in Figure 5.

Conclusion

The Quantitative Data Center is being developed in accordance with FAIR guidelines. Thus, to become a source for data exchange networks or part of an integrating platform or environment. Metadata and data accesses for automatic processing by programs according to common protocols in international standards.

At the current moment, the Center detected 700 data sets related to Russia from 400,000 processed records. It is also extracted 700 tables with quantitative information from 2000 scientific geological publications.

Further development of the Center is associated with increasing of data sources and adding features of geocoding for visualization of datasets on the map.

Acknowledgments. Scientific research is carried out within the State task of the State Geological Museum named after V. I. Vernadsky RAS on Topic No. 0140-2019-0005 “Development of an information environment for the integration of data from natural science museums and their processing services for Earth sciences”.

References

Brase, J. (2009), DataCite – A Global Registration Agency for Research Data, p. 257–261, Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology, Beijing, China. Crossref

Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 Crossref

DataCite Metadata Working Group.: DataCite Metadata Schema for the Publication and Citation of Research Data. Version 4.1. DataCite e.V. (2017) Crossref

Diepenbroek, M., H. Grobe, M. Reinke, et al. (2002), PANGAEA – an information system for environmental sciences, *Computers & Geosciences*, *28,* No. 10, 1201–1210, Crossref

Eremenko, V. S., V. V. Naumova, K. A. Platonov, et al. (2018), The main components of a distributed computational and analytical environment for the scientific study of geological systems, *Russ. J. Earth Sci.*, *18,* ES6003, Crossref

Emerson, C., E. M. Faustman, M. Mokrane, et al. (2015), *World Data System (WDS) Data Sharing Principles*, Zenodo, November 30, 2015. Crossref

Naumova, V. V., A. V. Belousov (2014), Digital repository "Geology of the Russian Far East" – an open access to the spatially distributed online scientific publications, *Russ. J. Earth Sci.*, *14,* ES1004, Crossref

Naumova, V. V., I. N. Goryachev, S. V. Dyakov, et al. (2015), Modern technologies of development of the Information infrastructure to support the research on geology of the Russian Far East, *Information Technology*, *21,* No. 7, 551–559. (in Russian)

Naumova, V. V., V. S. Eremenko, K. A. Platonov, et al. (2019), Development of geographically distributed information-analytical geological environment, *Russ. J. Earth Sci.*, *19,* ES6012, Crossref

Platonov, K. A. (2015), Methods and technologies for creation of the information processing system applied to publications on geology of the Russian Far East, *Russ. J. Earth Sci.*, *15,* ES4005, Crossref

Platonov, K. A., V. V. Naumova (2017), Methods and technologies for geological quantitative information integration, *Proceedings of Irkutsk State Technical University*, *21,* No. 2(121), 67–74, (In Russian)Crossref

Platonov, K. (2018), Methods and Technologies for Integration and Processing of Geographically Distributed Quantitative Geological Information, *Selected Papers of the XX International Conference on Data Analytics and Management in Data Intensive Domains* p. 250–255, CEUR Workshop Proceedings, Moscow, Russia.

Sarbas, B. (2008), The GEOROC Database as Part of a Growing Geoinformatics Network, p. 42–43, Geoinformatics 2008-Data to Knowledge, Potsdam, Germany.

Shokin, Yu. I., A. M. Fedotov, O. L. Zhizhimov (2015), Tehnologii sozdanija raspredelennyh informacionnyh sistem dlja podderzhki nauchnyh issledovanij, *Vychislitel'nye tehnologii*, *20,* No. 5, 251–274. (in Russian)

Wilkinson, M. D., M. Dumontier, I. Aalbersberg, et al. (2016), The FAIR Guiding Principles for scientific data management and stewardship, *Sci. Data*, *3,* 160018, Crossref

Corresponding author:
V. V. Naumova, V. I. Vernadsky State Geological Museum, Moscow, Russia. (naumova_new@mail.ru)