Development of geographically distributed information-analytical geological environment

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

Received 23 September 2019; accepted 25 November 2019; published 23 December 2019.

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. *KEYWORDS:* Information-analytical environment; integration of heterogeneous geographically distributed geological information; computational-analytical environment of geological information processing.

Citation: Naumova, V. V., V. S. Eremenko, K. A. Platonov, S. E. Dyakov, M. I. Patuk, and A. S. Eremenko (2019), Development of geographically distributed information-analytical geological environment, *Russ. J. Earth. Sci., 19,* ES6012, doi:10.2205/2019ES000696.

Introduction

The availability of developed information infrastructure is an essential condition of scientific researches provision effectiveness. Integration of information and computational resources in a unified environment and organization of access to them is one of the most important direction of modern information technologies. Global computer networks impetuous development leads to change of fundamental paradigms in data processing due to necessity of distributed information-computational resources support and development [*Shokin and Fedotov,* 2009; *Shokin et al.,* 2015].

Necessity of holistic integral information user field creation that consist of tools complex, analytical methods, geodata descriptions that are necessary for applied and fundamental researches in Earth sciences was noted in N. P. Laverov and co-authors' study [*Laverov et al.,* 2008]: "an approach to creation of holistic integral information field of user in Earth science implements an architecture of interconnected portals system, each of which is responsible for particular wide subject area, or individual aspects of the whole system technological functioning. The portals' system is connected by hyperlinks and represents a united distributed structure, the

[1]V. I. Vernadsky State geological museum of RAS, Moscow, Russia
[2]Institute of Automation and Control Processes of FEB RAS, Vladivostok, Russia

http://rjes.wdcb.ru/doi/2019ES000696-res.html

interaction in which is based on effective exchange of meta information and resources."

At the current moment to intensify scientific researches and development of science communications the access systems of open scientific publications, archives of science information, information systems and science museums data are developed.

The term "open access" was first mentioned on the Budapest conference on the open access in February 2002. Since that time its meaning was almost not changed: "Open Access" defines as free, immediate, permanent, full text, online access to scientific publications.

During 69th IFLA General conference at "Information technologies and Dublin Core metadata group work" seminar there were formulated principles. On these principles the ideology of "Opened archive" is based: consolidation within a world scale of science materials archives; free access to archives (to metadata); consensual archives and information providers interfaces; ease of use; application of current standards – HTTP, XML, Dublin Core, MARC, MARCXML.

With the development of Internet and related technologies, the tasks of science data accessibility, control, storage and propagation have reached a qualitatively new level. Multiple copying and publication of information on internet resources creates problems for its integration on the basis of common policies. The main difficulty is in the lack of identification mechanisms for the reliability of both data and their sources.

Domain names or IP-addresses doesn't have necessary stability for digital objects steady identification in digital space [*Green and Bide,* 1998]. To resolve this task in 1997 the DOI-system was created [*Paskin,* 2010]. The goal of DOI-names development is to uniquely identify a digital object and assign to it a permanent unique identifier. The developed technology based on Handle-system propose creation and storage of metadata in a format of DOI-entry for each digital object.

In 2009 work on the international project DataCite started [*Brase,* 2009]. The goal of the project – to provide direct access to scientific data through Internet. The project proposed a methodology of dataset citation and archiving method for further verification and reuse of research results in future. The project key feature – DOI registration agency foundation for scientific publication. Thus, research datasets are able to be registered using the DOI and become independent and unique objects.

The DataCite project is a partner of large scientific association CODATA (International Council for Science: Committee on Data for Science and Technology, http://www.codata.org/). The goal of this association is improvement of quality, reliability and accessibility of data and also the unification of efforts to manage, collect and exchange scientific data. One of the achievements of this community was a development of quantitative data citation principles in 2010–2014. The new approach offers a quantitative data publication procedure that includes assignment of DOI, creation of unified meta description and registration in a DataCite central repository (http://www.datacite.org). This registration is possible by allocating only in specialized systems with support of OAI metadata exchange protocols. According to DataCite portal data in 2016 it was registered more than 8.6 million of scientific datasets from 800 scientific and education organizations.

Integration of described approach with new type information system technologies is perspective, i.e. operations with both data and datasets. The storage object of such systems are datasets, i.e. data tables. The system receives new data through user interface or by exchange protocols of OAI metadata or data. To provide a uniqueness of each dataset the technology of DOI-names assignment is used. Thus, datasets with DOI-name can be definitely identified and have permanent link for Internet publication and scientific articles.

The following information systems can be cited as examples of data open access Systems and processing systems.

Digital Earth Australia (http://www.ga.gov.au/dea/home). Digital Earth Australia – is an Australia's government realization of a platform with open source code, developed within initiative of Open Data Cube (ODC). The DEA program presents code, documentation, users guide, tutorials and support for international users of Open Data Cube.

The Open Data Cube is a global initiative for increasing of abilities of satellite data usage. It provides users access to free and open data management technologies and analysis platforms. Application of free and open satellite data for ecological,

economic and social tasks can provide information and applications that makes a big impact on local, regional and global scales. Achievements in cloud calculations and availability of free and open technologies such as Open Data Cube means that developing countries without local infrastructure for big amount of satellite data processing can get access to data and processing power for creation of corresponding application and decision-making informing.

U.S. Geoscience Information Network (http://usgin.org). The main goal of USGIN: simplification of open access to modern digital data and applications in Earth sciences. USGIN standards, protocols and tasks – legacy of National Geothermal Data System (NGDS), collaborative data usage system that provides access to geothermal resources information.

Ausgin – Australian information network in Earth sciences (http://www.geoscience.gov.au). The system widely uses web-services – mainly web-cartographic services (WMS), but also a Web-Feature Services (WFS), Web Coverage Services (WCS).

OpenGeoscience BGS. British geological service has a wide spectrum of datasets. It constantly expands access to its services by publishing a large amount of data on the OpenGeoscience BGS portal (http://www.bgs.ac.uk/opengeoscien ce). OpenGeoscience is a free service where it is possible to view maps, download data, scan photos and other information. Services available in OpenGeoscience includes: geological data view through searching window of Great Britain geological map as well as using WMS; access to more than a million of scanned photos of geological sections and wells, and to photos from GeoScienic geological archive; view of published paper charts from 1832 to 2014 and publications from 1835 till now.

Portal EarthChem, supported by Columbia university (http://www.earthchem.org). This resource contains more than 860 thousand samples from 20 thousand scientific geological publications and provides the ability of analysis and visualization of geochemical databases' content on the map, such as GeoROck, PeDB, CedDB and others.

GEOROCK portal (http://georoc.mpch-mainz.gwdg.de/georoc). GEOROC (Geochemistry of Rocks of the Oceans and Continents) is supported by Max Plank Institute of Chemistry in Mainz. The database represents a wide collection of published analysis of volcanic formations and mantle xenoliths. It contains basic and micro elemental concentrations, radiogenic and non-radiogenic isotope relationships, and analytical ages for whole formations, glasses, minerals and inclusions. Metadata includes in particular geographical location with latitude and longitude, a class and type of rocks formation, laboratory and reference materials and links.

In Russia the following information systems can be noted:

SOBR Rosnedra (https://sobr.geosys.ru). Purpose – information support of geological exploration of mineral resources and reproduction of the mineral resource base by regular updates of current and summary information. It also provides information relevance and interaction with industry information resources of Rosnedra system. System architecture represents a distributed information-communication complex of existing information resources and specially designed program-technological infrastructure. It provides mainstreaming and integration of different information systems' data, its automated search, processing and uniform presentation and central access based on Internet.

Geoportal IVS FEB RAS (http://geoportal.kscnet.ru) – thematic web-portal that provides a united access point to volcanology and seismology spatial data and services of Institute of Volcanology and Seismology of FEB RAS. The geoportal is an element of spatial data infrastructure of IVS FEB RAS. The purpose of the geoportal is an integration of wide complex of scientific information collected in the institute during many years of research, provision an ability of data search by their metadata and data interchange within a network, visualization of spatial data in a form of interactive maps. In accordance with the requirements for metadata, data and services interoperability the Geoportal is implemented on the basis of ISO standards and OGC (Open Geospatial Consortium) specifications.

In 2014–2017, the authors of the study carried out work on designing, implementation and testing of the first version of Internet-infrastructure for support and maintenance of scientific geological researches on the Far East of Russia [*Naumova et al.,* 2015]. As an approach basis for implementa-

tion of Infrastructure is a loosely coupled blocky infrastructure, based on differences in types of geological data: spatial, quantitative, bibliographic and based on experts' knowledge's. Every separate information block of Infrastructure for integration, storage and search of data applies different approaches and technological solutions. Resources integration stored in different information blocks of heterogeneous System is provided through united and unified access to them. It is most consistent with accepted international standards.

Development of Heterogeneous Geographically Distributed Information-Analytical Geological Environment

The main goal of the project is in organization of united access point to geological data of the whole territory of Russia and systems of its' processing using abilities of data search in geographically distributed processing-analytical nodes for data processing. The interaction with such nodes is based on web-services technology. Integration of polytypic geological data and processing services into united information-analytical environment on the basis of uniform politics will provide an ability of its complex analysis. It will let to obtain a qualitatively new knowledges about geological objects.

In 2018 at the "V. I. Vernadsky State Geological Museum of RAS" the works on the design and development of the Information-analytical environment for scientific research support in geology were started.

Main specifications of the Environment:

1. Information-analytical environment provides users with authentic geological information for use in scientific purposes.

2. Information sources – geographically distributed heterogeneous Internet-resources, the information in which is based on standardized metadata. The program solutions of such sources allow application of standardized protocols for its automatic integration into infrastructure being created and scientific materials of science organizations, libraries and data centers.

3. The portal represents a single access point to polytypic, geographically distributed geological information on the territory of Russia: to geological maps, to metadata of government geological reports, to cadaster of mineral deposits; to satellite data, scientific publications; to tables of quantitative data and others, and also to specialized services of its' analysis and processing.

4. The approach is based on loosely coupled block infrastructure. This infrastructure is based on differences in types of geological data: spatial, quantitative, bibliographic and based on experts' knowledges. In every separate information block of infrastructure for integration, storage and search of data different approaches and technological solutions are applied.

5. 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 generalized scheme of the Information-analytical geological environment is presented in Figure 1.

The scheme includes:

- Access through the Internet to different kinds of geological information: scientific publications, maps, satellite data, qualitative data, geological data from different databases and others;
- Ease of finding specialized data using thematic queries, as well as services for their processing;
- Search result visualization, including use of cartographic basis;
- Convenience in the geological data and services distribution at the level of data (metadata) that corresponds to international standards and protocols.

Information sources – geographically distributed Internet-resources, the information in which is based on standardized metadata, and program solutions that allow the use of standardized protocols for its automatic integration into infrastructure being created and scientific materials of scientific organizations, libraries, data centers et al.

Main information types:

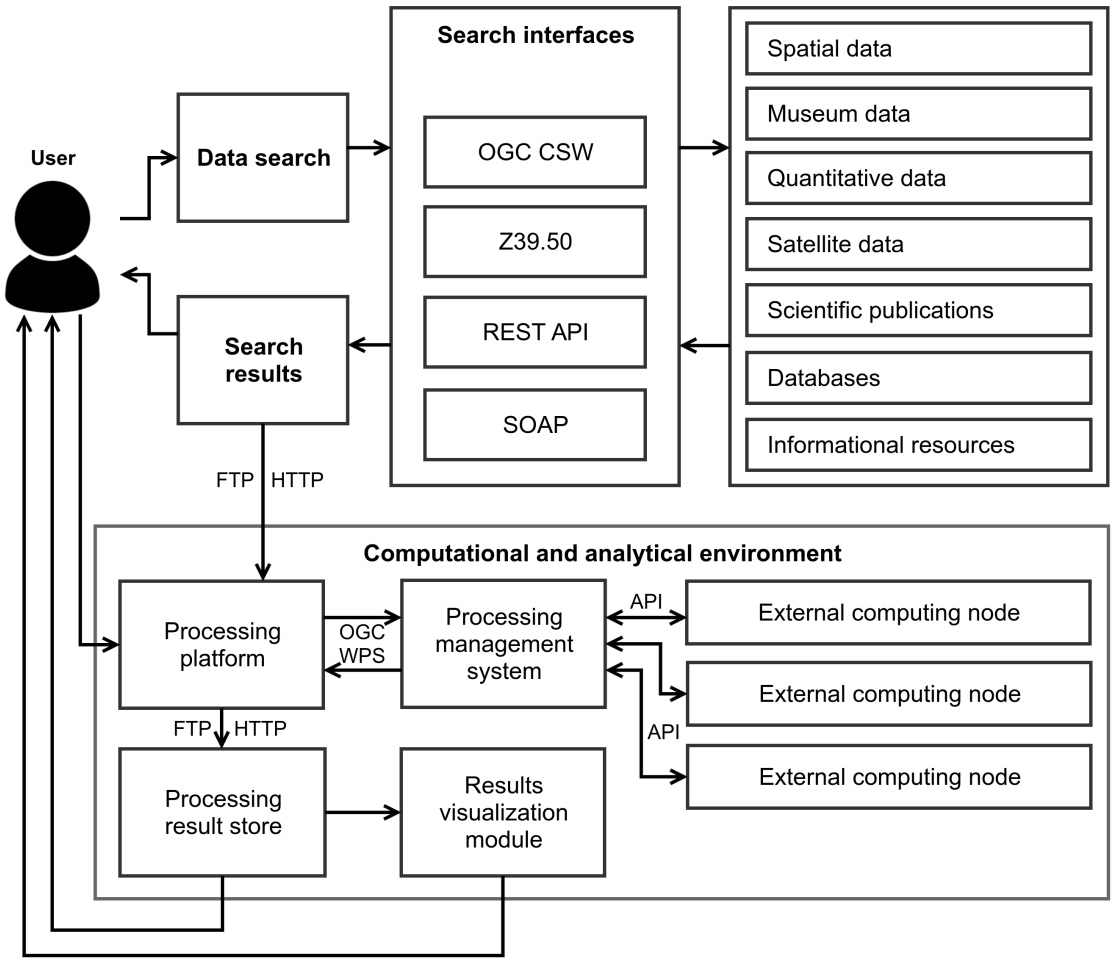

Figure 1. Generalized scheme of Information-analytical geological environment.

- Scientific publications: articles, monographs, manuscripts, dissertation abstracts and dissertations, tutorials and other materials;
- Cartographic information;
- Electronic information from Rosnedra DB;
- Satellite information from open sources – satellite monitoring centers providing accessible and reliable information;
- Quantitative information;
- Information of Natural Science Museums;
- Information materials about geologists and the history of geological study of the territory of Russia.

In this paper, there was no purpose to describe the legal side of data aggregation, although the authors know about it. The described System aggregates only those data that are distributed under OpenAccess conditions or under open licenses. It is open state information, information from digital repositories and world data centers, etc. All sources are referenced. For example, for cartographic information on VSEGEI website on the http://www.geolkarta.ru/info_proj.php page it is specified "The system allows to get free access to digital raster copies of sheets of a maps of scale $1 : 1,000,000$ and State maps of $1 : 200,000$ scale, as well as get the required sheets of high-resolution GGC on special request."

Block "Scientific publication" (http://geologyscience.ru/scientific_publications/) of infrastructure is an open access repository created on the basis of free access application DSpace, 6.3. Main its thematic – Earth sciences (geology, geochemistry, petrology, mineralogy, tectonics, geomorphol-

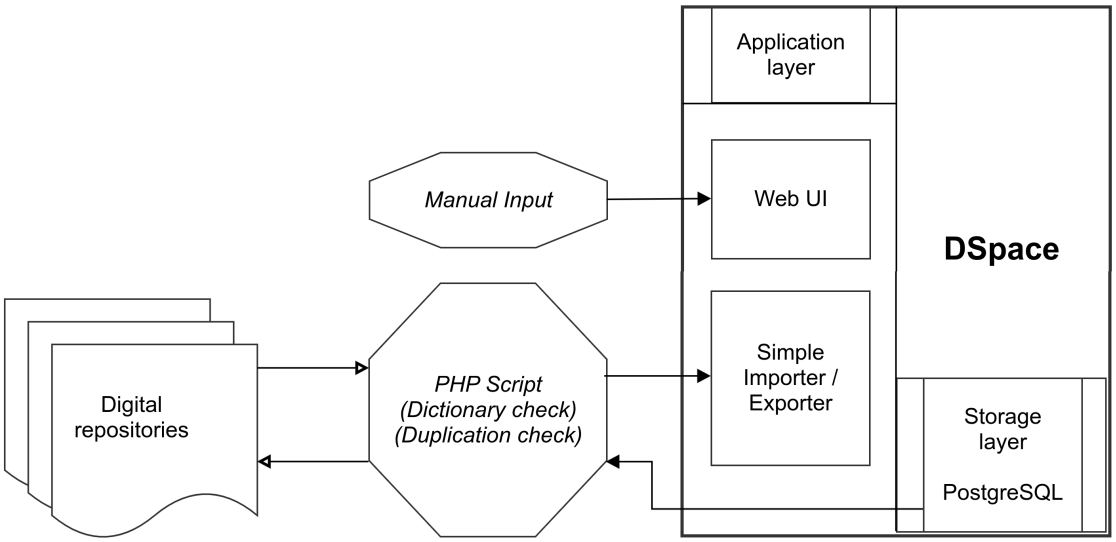

Figure 2. Generalized scheme of "Scientific publications" block.

ogy, volcanology, paleontology, stratigraphy and so on). The basis of the information are scientific articles, monographs, dissertations, dissertations' abstracts, reports' theses, conferences' materials in open access [*Naumova and Belousov,* 2014].

The PHP-script to search and extract the information from other repositories was created. The extracted information is filtered, i.e. it is automatically analyzed for concurrence with a dictionary of geological terms. The dictionary is created on the basis of keywords ∼2000 publications on repository thematic. The presence of 3 concurrences with dictionary is optimal. In this way it lets to select about 90% sources, corresponding to repository thematic. The other 10% are processed manually. This information is a basis for dictionary correction.

A large amount of information in the public access are texts in PDF format. Adding such data to the repository is impossible without the corresponding metadata. To extract metadata from such publications free software is used: Cermine – Context Extractor and Miner [*Tkaczyk,* 2015], FPDI – collection of PHP classes for PDF documents processing (https://www.setasign.com/products/fpdi/about/), PDFMiner – software for text information extraction from PDF based on Python (http://www.unixuser.org/∼euske/python/pdfminer/).

The information extracted from PDF files and other repositories is converted into SIP forman that is available for import into DSpace by standard tools.

The UDC (universal decimal classification) tag was added into DSpace to improve search of the information inside repository in addition to existed standard tools [*Naumova et al.,* 2015]. This information is extracted in half-automated mode from DSpace backup downloaded in a SIP format into text file with further upload by SQL-script into PostgreSQL table. Generalized functional scheme of repository is on the Figure 2.

The center of quantitative data (http://datacenter.geologyscience.ru/). The main source of quantitative information tables or datasets are scientific publications (Table 1). The system interacts with scientific publications' texts repositories using OAI protocol. Filtration of publications by subject is carried out on publication metadata. Quantitative data tables are detected and extracted from publications text and then are transformed into datasets [*Platonov,* 2018; *Platonov and Naumova,* 2017].

The PDF analysis method being developed should provide accurate retrieval of each character. Optical character recognition methods do not produce 100% result. Therefore, only "born-digital" materials are selected for processing. The PDF workflow involves adding optical character recognition methods as an additional step, but is not yet in use. These methods require the user to check the result of their work, which does not allow to automate the process, as in the case of "born-digital" materials.

Table 1. Sources of Geological Quantitative Information of RF

Sources (with example)	Data access	Data upload form	Metadata format	Storage form
Scientific publication repositories	HTTP OAI	Manually oai-pmh 2.0	oai_dc	Full text publications
"GeoRoc" databases	HTTP	Manually	Absent	Files with tables
FAIR "Pangaea" systems	API HTTP OAI	Manually	oai_dc pan_md datacite3	Metadata and datasets
DataCite network	API HTTP OAI	oai-pmh 2.0	oai_dc oai_datacite	Metadata

Institutions and universities created and maintained geological quantitative databases. Some databases are available from processing through Internet using RESTFull API. To get datasets the particular query with given constraints is formed. Metadata are formed from data extracted from result of the query.

The volume of a scientific article doesn't imply the publication of all initial information. The procedure according to the "principles of data citation" is applied to resolve the problem of accessibility of complete experimental data sets. Such systems-publishers are called World data centers. The Center of quantitative geological data of RF being developed by the authors in the same way collects metadata by OAI protocol, makes its filtration, stores selected metadata and organizes direct access to datasets.

DataCite network is a huge repository that collects and stores only metadata from all systems of scientific data management. When following the link where the metadata are published it is possible to find any type of information system: repository, database, world data center and so on.

General functional scheme of the Center of quantitative data is on the Figure 3.

The project provides the following parts of the Center for Quantitative Data: metadata collection and filtration blocks, service for extracting tables of quantitative data from scientific publications, block of metadata filling and extraction, storage system, search, cataloging, provision of datasets and metadata and processing node.

ISLANDORA (https://islandora.ca/) project software was used to organize the storage, provision and search of metadata. Metadata are located in Fedora repository and are available by OAI protocols. Search engine Solr provide metadata fields indexing and returns found records in JSON-format by requests through API.

Center's search system is divided into four semantic queries: "what?" – search by name, "where?" – spatial search, "when?" – temporary search, "who?" – search by personalities.

The system collects data in two languages: Russian and English. To perform search functions the Google machine translator and a module of Cyrillic transliteration into Latin were used. Thus, it became possible to search in Russian and discover geographic, geological and temporal terms for a cataloging system.

The Rosgeolfond' list of mineral deposits is used for cataloging. The dictionary of Russian geological objects is also used. Datasets binding is done using coordinates search or by geographical, geological and temporary terms in metadata.

The system uses two external computational nodes created by authors: the service for extraction of quantitative data tables from scientific publications and a node of quantitative tables processing. The architecture of nodes allows using them both for the tasks of the Center and for external API requests.

The Quantitative data center is being developed in compliance with FAIR recommendations. The task of data collection about Russia from foreign sources is being solved. The work on subject adaptation of Center's functions to understandable lan-

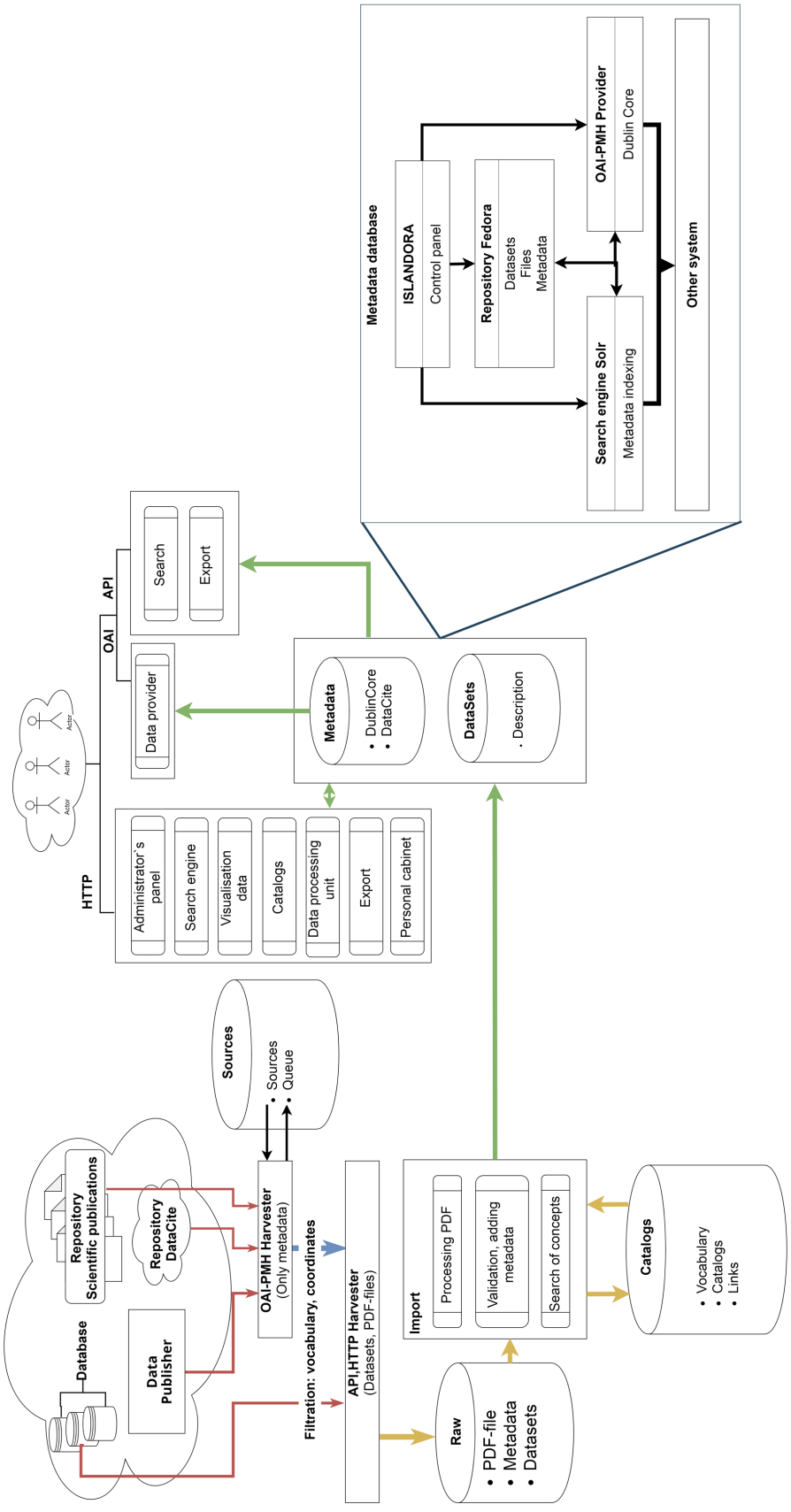

Figure 3. General functional scheme of the Center of quantitative data.

guage for geologist is being done: catalogs, search system, processing.

The Block of spatial data access (http://geologyscience.ru/spatial_information/) was developed to provide access to geological maps of Russia from geographically distributed heterogeneous sources. Access is implemented using spatial data metadata cataloging technology on the basis of Catalogue Service for the Web (OGC CSW) service catalogue international standard. Catalogue service provide an ability of fast data search using various criterion and receive attributive information about a particular object, including a link to the data. The use of metadata cataloging technology on the basis of international standards allows to integrate external data sources using this data provision approach. To describe geological maps metadata a profile of metadata on the basis of ISO 19115 and ISO 19139 standards is used. The data presented in the catalog is stored by the data provider on the external node as vector files and in a form of separate layers within spatial data access services such as OGC Web Map Service (OGC WMS) and OGC Web Feature Service (OGC WFS). To implement a catalogue service the open source program complex GeoNetwork is used. At the current moment the catalog contains metadata of the A. P. Karpinsky Russian Geological Research Institute (VSEGEI) on scale of $1 : 1,000,000$ of the 3rd generation across Russia (131 records) and also VSEGEI metadata on a scale of $1 : 200,000$ of the 2nd generation across the territory of Russia (212 records).

The block of OAO "Rosnedra" DB access (http://geologyscience.ru/bd_rosnedra/). In this block a remote access by request to deposits' metadata and statement geological reports being in DB "Rosnedra" is organized. This DB stores information about 52 thousand deposits and 478 thousand geological reports.

Deposit search uses faceted technology. User can select a deposit by name (a tooltip appears after entering first 4 symbols) or select an area, settlement etc. The type of minerals is selected separately (also on the basis of the tooltip). Such solution is a compromise between entering a full name (which is fraught with errors) and multi-page output of e.g. settlements. Thus, the search is performed by index, and not by a full-text. It provides to perform search with use of regular expressions. Search based on word forms in not supported.

Satellite block (http://sputnik.geologyscience.ru/) provides users with united access point to satellite data of Aqua, Terra, Landsat, Orbview-3 and other multispectral satellite data of high and middle resolution. The sources of this data are satellite data portals of FEB RAS, NASA, USGS. Data search is performed in one of three modes: search using data source tools, search using outer services metadata, search using own database of satellite images metadata [*Naumova and D'yakov, 2015*].

The MODIS radiometer data search and PDS-format data receiving is performed using https://oceancolor.gsfc.nasa.gov resource or from Satellite monitoring center of Institute of Automation and Control Processes of FEB RAS. Processing includes a calibration (products Lt, BT) and atmosphere correction (product rhos) and is performed with use of SeaDAS software. After that the creation of GeoTiff files and pseudo colored images using a library of functions, gdal program complex and Glance utility is performed (Figure 4).

The search of Landsat, ASTER/TERRA, EO-1, Sentinel 2A, 2B, Orbview-3 satellites' data is performed using database created on the basis of USGS metadata. This approach provides a high speed of search but leads to a delay in information about new images.

Satellite imagery is processed based on available information, but in any case, users get all the information, including preview images.

Landsat data is being calibrated, converted to projection files and are used to create pseudo colored images.

ASTER/TERRA data are including pseudo colored images of visible and IR channels and also data of physical measurements in HDF-files. ALI/ EO1 data partially covers the territory of Russia (significant gaps are present). They include georeferenced data (L1GST) in a format of GeoTIFF.

Sentinel-2 data are provided in a jpeg2 format together with georeferenced and calibration data and one large preview image in a GeoTIFF format. They include several images with different spatial resolution (Figure 5).

Panchromatic Orbview-3 satellite data are provided in a form of two GeoTiff files (with and with-

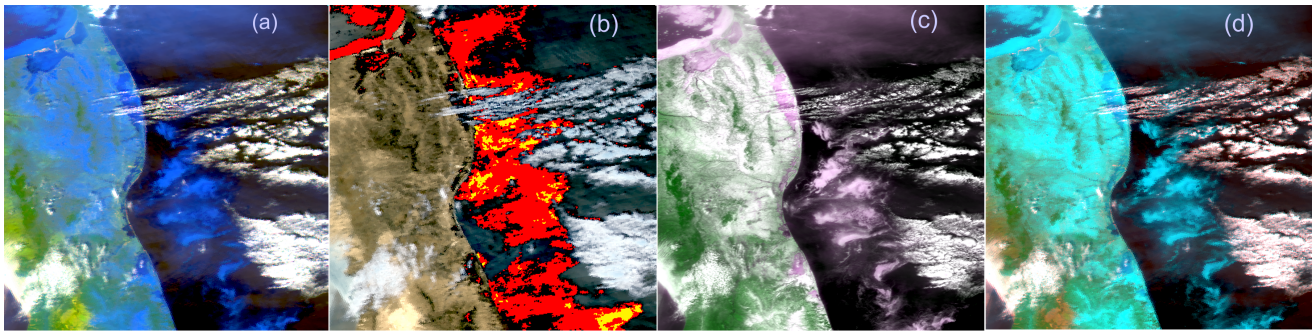

Figure 4. Pseudo colored images. Sakhalin coast. 2011-05-204. Combinations: (a) – 752, (b) – 23_22_20, (c) – 121, (d) – 721.

out a control points) and two thumbnail preview images.

The System of Portal data search (http://geologyscience.ru/) converts user request into a sequence of titles' searching queries (anuchinsky raion) and geographic coordinates search. After that queries are sequentially transferred to searching machines of Portal separate blocks (requests are processed by machines in a parallel mode). These searching machines use a various search method depending on the availability of the search interfaces of the blocks or direct access to the metadata databases. If any of the above tools are unavailable the global information systems with additional filtration of results are used for search. Queries for several searching machines are processed in a parallel mode.

The user sets the coordinates of the search area, and the system simultaneously access the services' search mechanisms. There are two problems with this: a) blocks are almost never support search by coordinate; b) different blocks have different search mechanisms with different performance (in this case the result should output in an individually).

To solve the problem (a) it was created a subsystem that receives a list of geographic objects, based on the specified territory (with use of OSM free service). It is impossible to get a full list of objects (a number of objects for a small settlement is more than thousands) and OSM service provide only a objects' hierarchy in a specified point (country, region, district). In this case a regular bypass of selected region with selection of 64 regular and 64 random points is performed into a list of objects. This objects' list is filtered so that only commonly appeared names of low levels of the hierarchy are used further. This solution is not optimal. The solution doesn't use an information about geological objects or even large geographic formations (such as Sayanskie gory) if they are not included into OSM database. To resolve the (b) problem the three level queries translator is used.

On the firsts level a query to every block is transformed into a set of individual queries. On the second level all individual queries are performed in a half-parallel mode (queries to different blocks are performed in a parallel mode, but there are no several parallel queries to each block). On the third level the queries' results are converted into a view adopted to output. Query maintenance is performed by a program working in a users' browser. As javascript is a single threaded language we use

Figure 5. Sample of Sentinel-2A image for 2018-05-09, Primorsky region.

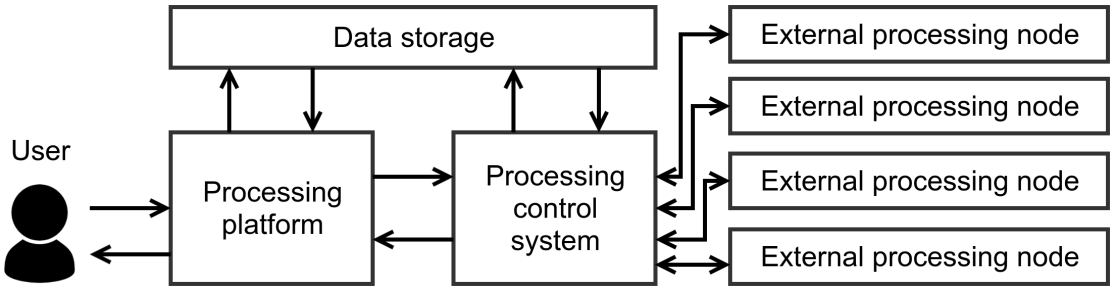

Figure 6. General scheme of information-analytical environment.

ordinary event processing model. Wherein first levels event processing is performed in a sequential mode, but individual queries are started in parallel mode through a set of executors. The results of executors' work go through a chain of handlers and are returned to users. This system has proven itself in terms of extensibility, flexibility and performance.

To process and analyze geological information the computing-analytical environment of geological data processing is being developed (http://service.geologyscience.ru). The Environment is a cloud instrument for users to process different types of geological data [*Eremenko and Naumova,* 2019; *Eremenko et al.,* 2018]. This environment provides to perform analysis of heterogeneous geological information, using external distributed and opened services of processing and analysis of various data types. Interaction with external services performs with use of intermediate interface of the service for launching processing procedures of spatial data of OGC Web Processing Service (OGC WPS). Each external service corresponds to separate WPS-process that is called when the selected service is accessed. Used WPS-processes are hosted within WPS service implemented on the basis of GeoServer open source program complex. The external services' metadata catalog for fast search and information receive provided by the Environment is developed. The catalog contains information about main services' functions, service provider and also a technical information including web-address, access protocols, a link to interface description and authorization mechanisms. The use of technical information from catalog allows to organize an external services' monitoring system to provide a high level of reliability during their usage. The general scheme of the environment is on the Figure 6.

At the current moment access to following processing nodes is implemented: multidimensional methods of data analysis (SMG RAS), structural analysis of scientific publications (Warsaw University), natural language processing (Sheffield University), petrological and geochemical data processing (IFZ RAS), table data visualization (Plotly project). All external processing and analytical services are included into services' catalog. The catalog is developed for convenient search and obtain of the information about particular service. To provide a high level of analyzing and processing services' work reliability the environment status and standalone service by itself monitoring system was developed.

At the current moment the Computational and analytical block includes following processing nodes:

- Multidimensional methods of data analysis. It includes a set of methods for multidimensional analysis of quantitative data such as factor analysis, cluster analysis, regression analysis and so on. As a component for module implementation of quantitative data statistical analysis it was a R programming language selected.
- Satellite data processing. It includes methods of primary satellite data processing such as calibration and satellite data image navigation.
- Petrological and geochemical data processing. The interactive database of petrological and geochemical data processing methods was developed in Institute of Physics of the Earth of RAS [*Ivanov,* 2016]. The system provides services of spidergrams, histograms and classification charts creation; services of minerals identification on the basis of their chemical composition; service of mineral composition interpretation and decomposition into mynals

and so on. The interaction interface is based on the REST architecture.

- Structural publication analysis. A service has been developed at the multidisciplinary center for mathematical and computational modeling (University of Warsaw, Poland) for scientific publications metadata extraction [*Tkaczyk,* 2015]. Metadata includes authors, affiliation, abstract, keywords, magazine name, volume, year of issue, parsed bibliographic references, section structure of the document, section headings and paragraphs. The interaction interface is based on the REST architecture.
- Natural language processing. At the University of Sheffield within the GATE project (General Architecture for Text Engineering) a number of textual data processing services for various language types have been developed [*Maynard et al.,* 2016]. For textual data processing in Russian services to determine the parts of speech of words, as well as the allocation of named entities, such as names and surnames, organizations' titles, geographical names, dates, monetary units, etc. The interaction interface is based on the REST architecture.

Monitoring the availability of resources when using data and services from external information systems is an important part of the information and analytical environment described. Currently, an approach for monitoring external processing services has been developed and implemented. This approach involves checking the availability of the node hosting the service, checking the availability of the service using the specified interaction protocol, and checking the service for changes in operation using test requests. In the future, the developed monitoring system will be used to monitor all external information systems used in the information and analytical environment.

If you do not have access to data or services from an external information system, the user is notified when you attempt to use the data or services of the selected information system. If protocols or access interfaces are changed, the environment administrator is notified. Duplication of data from external information systems is not implied within the approach used.

Conclusion

It is the first time for the information-analytical support of scientific researches in geology the following has been done:

- Creation of information-analytical environment model in accordance with stated requirements. The model was a basis for the environment architecture with components' description and links between them.
- Designing of information-analytical environment on the basis of analysis of international and domestic experience of geographically distributed systems creation and development;
- The methods and technologies for aggregation, monitoring, storage, processing and analysis of geographically distributed geological information and its processing services are developed and adapted;
- The development of information-analytical environment for heterogeneous geographically distributed geological information and its processing services was implemented.
- The Portal beta-version is accessible at the Internet (http://geologyscience.ru/). Preliminary tests are currently underway. Trial operation was organized to test the capabilities of providing access to geographically distributed scientific geological resources and its processing and analysis services.

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

References

Brase, J. (2009), DataCite – a global registration agency for research data, *Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology (COINFO 09)* p. 257–261, IEEE, Los Alamitos. **Crossref**

Eremenko, V. S., V. V. Naumova (2019), Sistema katalogizacii i monitoringa territorial'no raspredelennyh vychislitel'nyh uzlov v srede WPS servisov dlja reshenija geologicheskih zadach, *Vestnik NGU.*

Seriya: Informacionnye Tehnologii, *17,* No. 2, 39–48. (in Russian)

Eremenko, V. S., V. V. Naumova, K. A. Platonov, S. E. Dyakov, A. S. Eremenko (2018), The main components of a distributed computational and analytical environment for the scientific study of geological systems, *Russian Journal of Earth Sciences*, *18,* No. 6, ES6003, **Crossref**

Green, B., M. Bide (1998), *Unique Identifiers: A Brief Introduction*, 11 pp. Book Industry Communication/EDItEUR, London.

Ivanov, S. D. (2016), Interaktivnyj reestr geosensorov na osnove veb-prilozheniya, *Komp'yuternye Issled. i Modelirovanie*, *8,* No. 4, 621–632. (in Russian)

Laverov, N. P., Yu. M. Arskyi, et al. (2008), An integrated information field in earth sciences, *Herald of the Russian Academy of Sciences*, *78,* No. 5, 439–442, **Crossref**

Maynard, D., K. Bontcheva, I. Augenstein (2016), Natural Language Processing for the Semantic Web, *Synthesis Lectures on the Semantic Web: Theory and Technology*, *6,* No. 2, 194, **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, *Russian Journal of Earth Sciences*, *14,* No. 1, 1–8, **Crossref**

Naumova, V. V., S. E. D'yakov (2015), Organizacija dostupa i analiz dannyh distancionnogo zondirovanija dlja nauchnyh issledovanij v geologii na Dal'nem Vostoke Rossii, *Vestnik KRAUNC, Nauki o Zemle*, *1,* No. 25, 60–65. (in Russian)

Naumova, V. V., I. N. Goryachev, S. E. D'yakov, A. V. Belousov, K. A. Platonov (2015), Sovremennye tekhnologii formirovaniya informacionnoj infrastruktury dlya podderzhki i soprovozhdeniya nauchnyh geologicheskih issledovanij na Dal'nem Vostoke Rossii, *Informacionnye Tekhnologii*, No. 7-S, 551–559. (in Russian)

Paskin, N. (2010), "Digital Object Identifier (DOI®) System", *Encyclopedia of Library and Information Sciences* p. 1586–1592, CRC Press, Boca Raton.

Platonov, K. A., V. V. Naumova (2017), Metody i tehnologii integracii kolichestvennoj informacii v geologii, *Vestnik IrGTU*, *21,* No. 2 (121), 67–74. (in Rusian)

Platonov, K. A. (2018), Methods and Technologies for Integration and Processing of Geographically Distributed Quantitative Geological Information, *DAMDID/RCDL 2018 (Moscow, Russia, October 9–12, 2018), CEUR Workshop Proceedings, vol. 2277, Selected Papers of the XX International Conference on Data Analytics and Management in Data Intensive Domains (L. Kalinichenko et al., eds.)* p. 250–255, CEUR, Moscow.

Shokin, I. Yu., A. M. Fedotov (2009), K voprosu o razvitii informacionnoj infrastruktury SO RAN, *Vychislitel'nye Tehnologii*, *14,* No. 6, 127–137. (in Russian)

Shokin, I. Yu., A. M. Fedotov, et al. (2015), Jevoljucija informacionnyh sistem: ot web-sajtov do sistem upravlenija informacionnymi resursami, *Vestnik Novosibirskogo Gos. Universiteta. Seriya: Informacionnye Tekhnologii*, *13,* No. 1, 117–134. (in Russian)

Tkaczyk, D., P. Szostek, et al. (2015), CERMINE: automatic extraction of structured metadata from scientific literature, *International Journal on Document Analysis and Recognition*, *18,* No. 4, 317–335, **Crossref**

Corresponding author:
V. V. Naumova, Vernadsky State Geological Museum, Mokhovaya St. 11, bldg. 11, 125009 Moscow, Russia. (naumova_new@mail.ru)