Information and Analytical Environment to Support Scientific Research in Geology: Current Status and Development Perspectives

Kirill A. Platonov
Vernadsky State Geological Museum RAS
Moscow, Russia
k.platonov@sgm.ru

Vera V. Naumova
Vernadsky State Geological Museum RAS
Moscow, Russia
v.naumova@sgm.ru

Vitaliy S. Eremenko
Vernadsky State Geological Museum RAS
Moscow, Russia
vitaer@gmail.com

Sergey E. Dyakov
Institute of Automation and Control Processes FEB RAS
Vladivostok, Russia
sergdkv@gmail.com

Abstract

The work describes the development and adaptation of methods and technologies for processing and analysis of territorially distributed heterogeneous geological information and services of its processing.

The basis of the information-analytical environment for geology scientific support and maintenance, designed on the base of created approaches, methods and technologies, integrating territorially distributed geological information with the use of special services, its analysis and processing.

The authors suppose that the developed platform of processing and analysis thematic services management, which is a part of the information-analytical environment, will provide an access to the modern knowledge-intensive algorithms and computing resources storages, required for expeditious processing of geological data large arrays.

1 Open Access Systems to Data and Processing Systems

The term “open access” was for the first time mentioned at the Budapest conference on open access in February 2002. The meaning is not practically changed: Open Access is a free, immediate, permanent, fulltext, online access to scientific information.

The seminar “Information Technologies and Dublin Core Metadata Group operation” within the frames of 69th IFLA General Conference defines the main principals of “Open Archive” ideology: world-wide consolidation of scientific materials archives; open access to archives (metadata); consistent archives and information providers interfaces; usage simplicity; application of existing standards - HTTP, XML, Dublin Core, MARC, MARCXML.

The open access systems to scientific publications, to scientific information archives, natural history museums data etc. are developed in the present moment for intensification of scientific research and scientific communication development.

The actual task is to provide open access to information and digital data and also to Earth sciences data analysis and processing systems.

The most known systems at present are:

Digital Earth Australia (http://www.ga.gov.au/dea/home) - the Australian government platform for analysis with open source code, developed within the frames of Open Data Cube (ODC) initiative. DEA Program provides the code, documentation, manuals, textbooks and support for Open Data Cube international users. The open data cube (ODC) is the global initiative for satellite data application possibilities increase. It provides users

with the access to free and open data management technologies and analysis platforms. The application of free and open satellite data for ecological, economic and social purposes is able to provide information and applications, highly influencing local, regional and global scales. Achievements in the cloud computing and availability of free and open technologies, such as Open Data Cube, mean that different countries without local infrastructure for data large volumes processing are able to receive access to data and computing facilities for creation of corresponding applications and decision making information.

The main *U.S. Geoscience Information Network* (http://usgin.org) target is to simplify open access to united digital data and software in Earth sciences. USGIN standards, protocols and tasks is an inheritance of National Geothermal Data System (NGDS), the system of mutual data usage, providing the access to geothermal resources information.

British Geological Survey has a wide range of data sets and permanently expands the access to them by publication of data large quantities on the *OpenGeoscience BGS* (http://www.bgs.ac.uk/opengeoscience) Portal. OpenGeoscience available services include: viewing geological data through the UK geological map search box, and by using WMS, the access to more than million geological sections and pits photo scans, and also to GeoScienic geological photo archive; viewing of published paper maps from 1832 to 2014 and publications from 1835 upto the present moment.

EarthChem Portal, supported by Columbia University (http://www.earthchem.org) contains more than 860 thousands samples from 20 thousand geological publications and provides possibility of GeoRock, PetDB, CedDB etc. geophysical data bases analysis and visualization on the map.

2 Vernadsky State Geological Museum RAS (SGM RAS) Project "Development of information-analytical geological environment for geological researches support GeologyScience.ru»

In 2014-2017 the authors executed works on development and realization of first Internet-infrastructure version for support and maintenance of geological scientific researches in the Far East of Russia. [Nau15; Nau17].

The main target of that project was organization of the single access point to geological data on the Russian territory and to the processing systems with the possibility of data search in territorially distributed diverse sources and also with use of territorially distributed computing-analytical data processing knots; the interaction with them is through web-services technologies. The integration of heterogeneous geological data and processing services into united information-analytical environment based on united policies provides possibility of complex information analysis and allows to receive a qualitative new knowledge about geological objects.

The loosely coupled block infrastructure is on the base of suggested approach. This infrastructure based on difference in geological data types: spatial, quantitative, bibliographic and based on expert knowledge. Each separate information environment block has different approaches and technologies for data integration, storage and search.

The main Environment structure requirements are:

- Access to information resources on the base of international standards and unified policies;
- Pass-through information search in the Environment as on the logical, as also on the physical level;
- Monitoring of territorially distributed data sources and computing knots, and also the Environment main knots.
- Support of pass-through authorization and rights differentiation.

Figure one represents a generalized scheme of Information-analytical geological environment. The Environment contains 2 basic levels: information and computing (Figure 1)

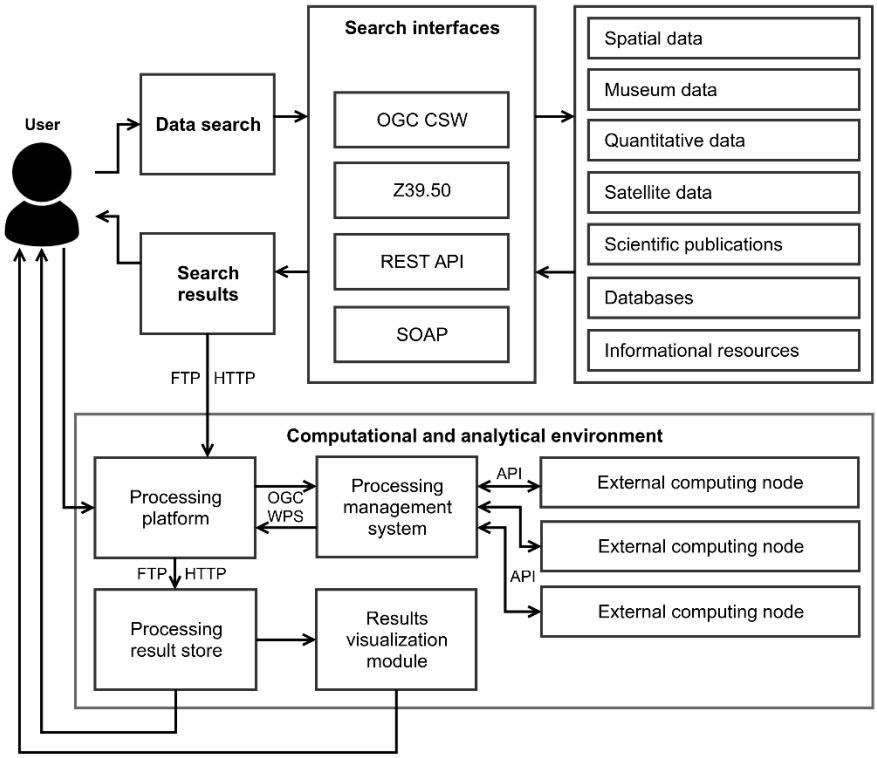

Figure 1: Generalized functional scheme of Information-analytical geological environment.

3 Environment information level

Unification of thematic resources into one integrated information infrastructure of scientific researches support allows to receive strict access to all System knots, decreases user's operations with each knot. It allows to separate knots and services to integrate into other information systems and simplifies System administration process.

The information resources integration requires the following functions [Sho09, Sho15]:

1. Access to each integrated resource through united user interfaces on unified protocols;
2. Pass-through search in the whole set of integrated information resources and also in their logical and physical subsets;
3. Information extraction in unified formats;
4. Resources and access management in accordance with unified policies;
5. Integrity and accessibility control for all resources services;
6. Collection of the statistic information about resources usage.

Information resources are territorially distributed Internet-resources, the information in which is based on standardized metadata and their program decisions allow standardized protocols application for its automatic integration into developed infrastructure, and also scientific organizations, libraries, data centers etc. materials.

We put in words the main requirements to Environment information blocks:

- Data uniformity within the block;
- Information extraction from territorially distributes sources is to be in unified formats for each block;
- Metadata DB is to be in the given data type standards;
- API search.

"Scientific publications" block is an open access repository, based on DSpace, 6.3. The information base is scientific articles, monographies, theses, author's abstracts, report abstracts, conference proceedings.

The script on PHP language was created for search and extraction of information out of other repositories. The extracted information is filtered, i.e. automatically analyzed for identity with the geological terms thesaurus. Thesaurus is created on the base of key words ~ 2000 publications on repository themes. The optimum is 3 matches with the thesaurus. It is possible to select more than 90% sources corresponding to repository themes. The rest 10% is for hand processing. This information is for thesaurus correction.

Texts in PDF format are a huge part of open access information. To extract metadata from such publications, free software is used: Cermine- Content Extractor and Miner, FPDI – PHP classes collection for PDF documents processing, PDFMiner – software on Python for text information extraction out of PDF.

The extracted from PDF files and other repositories information is to transform SIP format available for import to DSpace by standard means. For improvement of information search in repository, it is added UDS (universal decimal system) to existing standard DSpace search tags. This information is extracted in semi-

automatic mode out of DSpace uploaded backups in SIP formats into text file with the further SQL script download into PostgreSQL DSpace table.

Quantitative data block. The developing block is a single access point to geological territorially distributed quantitative information through unified interface [Pla18]. The data are integrated on logical level with the use of global DataCite metadata scheme. Scientific publications, author's data bases, world data Centers, and also experimental data tables are chosen as data sources. The definition and data receipt in information system is possible on OAI protocols and RESTfull interface. Integrated information, if necessary, is transformed by author's algorithms into electronic tables' format and metadata are generated.

Spatial data access block is developed to provide user access to Russian geological maps. An access is achieved using spatial data metadata cataloging technology on the base of international service of OGC Catalogue Service for the Web (OGC CSW) catalogue. Catalogue service allows a quick data search on different criterions and to receive attributive information about separate object including data reference. Metadata cataloging technology application on the base of international standards allows to integrate external data sources, using this data representation approach. Metadata profile on ISO 19115 and ISO 19139 standards are used for geological maps metadata description. Catalogue data are stored at data supplier on the external knot in the form of vector files and also in the form of separate layers in the frames of spatial data access services, such as OGC Web Map Service, (OGC WMS) and OGC Web Feature Service (OGC WFS). A software package with the open code GeoNetwork is used for catalogue service. The catalogue at the moment has metadata of A.P. Karpinsky Russian Geological Research Institute (VSEGEI) 1:1000000 scale third generation on the Russian territory (131 records), and also VSEGEI metadata 1:200000 of the second generation on the Russian territory (212 records).

Remote access to Meta information about Russian Federation mineral deposits and state geological reports of Rosnedra agency data base is fixed in the Environment, containing records about 52 thousand deposits and 478 thousand geological reports. Facet technology is used for deposits search. A user can select a deposit on name (after first four symbols a hint arises), or select region, town etc. The minerals type (on the base of arising hint) is to be selected separately. This decision is a compromise between input of full names and output of multipage list. Thus, the search is a selection not by index, but by fulltext search, allowing it with regular phrases use.

Satellite block provides to users a single access point to спутников Aqua, Terra, Landsat, orbview-3 satellite data and to other multispectral data of high and medium resolution. The data sources are satellite data portals of Institute of Automation and Control Processes FEB RAS Satellite Monitoring Center, NASA, USA Geological Survey (USGS). The data search is in three modes: search with user's external search portal, search on metadata, search on own metadata satellite images data base.

Satellite images are processed according to existing information, but in any case a user receives the whole information, including overview images.

Search system converts a request generated by user into requests of names search sequence and search on geographical coordinates. After that, the requests sequentially from user's browser are transferred to search machines of the Portal separate blocks (machines process requests in parallel), and they, in turn, searching independently or applying to their own search systems of the Portal blocks or to global search systems.

Integration of the described approach with the technologies of information systems new type, i.e. operations with data and also with data sets, is very perspective. The object of these systems storage is data set, i.e. tables. New data come in the system through user interface or on metadata exchange protocols and OAI dada.

Computing-analytical Environment block is a cloud user's tool for different types of geological data processing. The suggested approach supposes an application of external computing knots for data processing, interaction with those, are realized through web service technologies use, in particular OGC Web Processing Service.

The implemented platform is an intermediary between the user and external processing systems, providing a single interface of access to all processing algorithms, existing in external processing systems (system knots). The described architecture also contemplates data usage possibility, not only out of available in open sources, but also download of data for processing by user himself. The developed Environment computional - analytical processing blocks and geological information analysis are organized in the form of service and analytical functions sets with the possibility of user's access to processing method selection; processing chains, switching on data download, formats transformation, analysis methods and results visualization; thematic chains implementing a sequence of analysis methods. The access to processing and analysis services is available through the platform of data processing distributed services management.

Computing-analytical geological environment [Ere18] at the present moment includes the following processing knots:

- 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 etc. As a component for module of statistical analysis of quantitative data was selected the R programming language. The interface for interaction with services is organized with module Rserve usage. This knot is developed and supported by Vernadsky State Geological Museum RAS [Pla18].

- Satellites data processing. It includes methods of satellite data initial processing, such as calibration and satellite data spatial binding. This knot is developed and supported by Institute of Automation and Control Processes FEB RAS.
- Petrology-geochemical data processing. The Shmidt Institute of Physics of the Earth RAS developed an interactive base of petrology-geochemical data processing methods [Iva16]. The system represents services of spidergrams, histograms and classification diagrams design; the service of minerals identification on their chemical composition; minerals composition interpretation service and decomposition into minals etc. The interface of interaction with services is built on the base of REST architecture.
- Publications structural analysis. Interdisciplinary center of mathematical and computing modelling (Warsaw University, Poland) developed a service for metadata extraction out of scientific publications [Tka15]. Metadata include authors, affiliation, abstract, key words, journal name, volume, year of issue, sorted bibliographical references, structure of document chapters, chapters names and paragraphs. The interface of interaction with services is built on the base of REST architecture.
- Natural language processing. Sheffield University in the frames of GATE project (General Architecture for Text Engineering) developed a range of text data processing services for different languages [May16]. For text data processing in Russian language, services are provided on words parts of speech definition, and also for allocation of named entities such as names and surnames, organizations names, geographical names, dates, monetary units etc. The interface of interaction with services is built on the base of REST architecture.

Monitoring system is developed in the frames of computing-analytical environment for high level of services operation, allowing to operatively reply on services operation changes [Ere19]. Applications of heterogeneous services, interaction with those are executed with the help of different protocols and on different interfaces, implicate a list of monitoring restrictions. However, having united technical information about each service (service web-address, protocol, protocol version etc., it is possible to implement the following types of verifications:

- Test of remote knot access;
- Test of remote knot service functionality on required interaction protocol;
- Test of service operation alterations on the base of WPS processes requests.

The more complicated service conditions tests types require detailed description of interaction interface, and such tests are dependent on concreate service realization.

Using described service conditions test methods, it is possible to form statistic of separate services accessibility. In case of any access problem to concrete service, it is possible to offer users alternative realizations of such service if available in the environment.

4 Development perspectives

Thematic resources unification into a single integrated information infrastructure of scientific researches support allows to receive a direct access to all System knots. It decreases user's operations with each knot, allows to separate knots and knots separate services, to be integrated with the other information systems, simplifies System management process.

To solve a range of problems: creation of united metadata Data Base, development of simple , development of simple general search tool to each information block and search results transfer possibilities, the authors, at the present moment are developing an approach allowing integration of heterogeneous metadata bases out of Environment information blocks into united subsystem on CKAN platform. CKAN is a powerful system of data management with the open program code, making data accessible, providing data optimization tools, their mutual application, allocation, representation and storage. CKAN Platform is a new type of information systems – data management systems (DMC), based on "open access" principals and CODATA operations. The storage object of this system is data sets i.e. tables. New data appear in the system through user interface or through metadata exchange protocols and OAI data.

References

[Ere19] Eremenko V. S., Naumova V. V. System of cataloguing and monitoring of geographically distributed computing nodes in the environment of WPS services for solving geological problems // NSU Bulletin. Series: Information technologies. 2019. T. 17, No 2. Pages 39–48. DOI 10.25205/1818-7900-2019-17-2-39-48

[Iva16] Ivanov S.D. Online register of geosensors based on web application//Computer research and modeling, 2016. T. 8. No. 4. Pages 621-632.

[Nau15] Naumova V. V., Goryachev I.N, Diakov S.E., Belousov A.V., Platonov K.A. Modern technologies of information infrastructure formation to support and support scientific geological research in the Far East of Russia//Information technologies, 2015, № 7 - Pages 551-559

[Nau17] Naumova V. V. Information and Functional Capabilities of Information Internet Infrastructure to Support Scientific Research in Geology // XVI Russian Conference "Distributed Information and Computing Resources. Science - Digital Economy "(DICR-2017): Reports of the XVI All-Russian Conference (December 4-7, 2017). Novosibirsk / Under Ed. O.L. Zhizimov, A.M. Fedotov. - 2017. - Novosibirsk: ICT SB RAS. – Pages 44-49 - ISBN: 978-5-905569-10-4.

[Sho09] Shokin Yu.I., Fedotov A.M. To the issue of development of information infrastructure of the SB of Russian Academy of Sciences//Computing technologies. - 2009. - T.14. - № 6. - Pages 127-137

[Sho15] Shokin Yu. I., Fedotov A. M., Zhizimov O. L. Technologies of creating distributed information systems to support scientific research//Computing technologies - 2015. -T. 20, № 5. - Pages 251-274

[Tka15] 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, Pages 317-335, doi: 10.1007/s10032-015-0249-8.

[May16] 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

[Ere18] 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

[Pla18] Platonov K. 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, eds. Leonid Kalinichenko, Yannis Manolopoulos, Sergey Stupnikov, Nikolay Skvortsov, Vladimir Sukhomlin, Pages 250–255.