RUSSIAN JOURNAL OF EARTH SCIENCES
vol. 18, 6, 2018, doi: 10.2205/2018ES000636

The main components of a distributed computational and analytical environment for the scientific study of geological systems

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

[1]Vernadsky State Geological Museum RAS, Moscow, Russia

[2]Institute of Automation and Control Processes, FEB RAS, Vladivostok, Russia

Abstract. The article describes the way to organize the processing and analysis of the geological data using a distributed computational and analytical environment. We propose various types of geological data used by researchers, as well as some methods for their processing and analysis. Also, we present a model of the computational and analytical environment for processing and analysis of the geological data. We constructed the environment architecture and described the requirements for each component of this environment, based on the proposed model. The proposed approach requires the use of external data processing computing nodes. Interaction with the nodes is carried out using web services technology, in particular OGC Web Processing Service. The study describes the capability of some computational nodes for processing inconsistent geological data and provides

This is the e-book version of the article, published in Russian Journal of Earth Sciences (doi:10.2205/2018ES000636). It is generated from the original source file using LaTeX's **ebook.cls** class.

examples of methods used for working with different types of data. For the comprehensive solution of processing tasks, we propose to use processing chains that allow several processes to be executed both sequentially and in parallel. The results of one or several processes can be transferred as a source for another process. We provide an access to services for processing and analyzing geological data both through the web interface and through the application programming interface to interact with external information systems, with the ability to download data directly from the user's computer, or by specifying a direct link to data from an external resource.

Introduction

It is the pressing challenge to provide geological research with tools for data processing and analysis on the Internet. Researchers are interested in the possibility of processing and analysis of various types of data remotely using modern means.

Processing and analyzing data in geology requires the application of a large number of different algorithms, processing procedures and corresponding software solutions. Often, when researching a particular object in geology, it is necessary to use various types of data, such as geological maps, quantitative data, satellite data, museum data, text data, etc.

Each type of geological data has its own set of analysis and processing algorithms with the appropriate software that implements mentioned procedures. When analyzing the quantitative data, the following modern data analysis methods are used: regression analysis, principal component analysis, hierarchical cluster analysis, k-means, factor analysis, linear discriminant analysis, canonical variables analysis, etc. While analyzing the spatial data, various methods of cartographic analysis are applied. Operations can be either simple (for example, geometric operations on vector data), or very complex, up to calculation of global models. When processing satellite data, atmospheric correction methods are applied, as well as conversion of coordinate systems to the GIS user coordinate system; consolidation of satellite images, etc. Textual analysis of scientific publications implies methods of automatic extraction of various parts of publications (metadata, tables, charts, images, reference lists, etc.), as well as thematic concepts and various methods of creating ontologies.

Although it is a difficult task to conduct an integrated analysis of different types of data while solving a particular problem. Therefore, it is necessary to provide researchers with an opportunity to analyze and process heterogeneous geological data within a single system; it allows to integrate the obtained results into a qualitative outcome. Due to the continuous growth in the amount of geological data and its processing tools, a more promising approach is to use the remote computing systems that provide processing services for certain types of geological data.

Analysis of the Current Research in This Field

Currently, developers of data processing software solutions are actively implementing cloud computing into the processing process. Usage of external services instead of user applications allows to process data on the most suitable equipment. Thus, data processing becomes more efficient, and the user is able to process data with the most up to date algorithms using the web interface, without the need to install, set up and support processing software on his personal computer.

Development of data processing information systems based on interaction with the user through the web interface is actively conducted in various subject areas. Thus, in the field of satellite data processing such systems include the NASA Giovanni project, the Google Earth Engine system, ESA G-POD platforms, as well as the See the Sea, Vega-Science, etc. information systems created in Space Research Institute RAS based on GEOSMIS technology [*Tolpin et al.,* 2011]. As for the spatial data processing and analysis, an example of such an information system is ArcGIS Online.

These information systems are aimed at working with a certain type of data, such as satellite, spatial, quantitative or textual data. However, the use of one or several systems to solve a number of geological research problems is extremely difficult because of the lack of mechanisms for interaction between heterogeneous systems. To work with heterogeneous geological data, it is necessary to use an approach that allows the interaction of various systems and processing services in a single computational and analytical environment. Such an approach to organize user interaction with various systems of data analysis and processing was proposed by the ICT SB RAS employees [*Fedotov et al.,* 2006]. According to this approach the interaction interface between the user and data processing services uses the technology of web-services, in particular, OGC Web Processing Service interface. Considering the peculiarities of working with geological data, this approach can be adapted to create a computational and analytical environment that allows to process geological data using external analysis and data processing services.

Purpose and Objectives

The purpose of this work is to create a computational and analytical environment to process geological data using external analysis and processing services. It is necessary to create a computational and analytical environment model in accordance with the proposed requirements. On its basis the general structure of the environment will be created, with the description of components and connections between them.

Designing a Computational and Analytical Environment for the Solution of Geological Problems

In the proposed model of the computational and analytical environment in geology, we distinguish four main objects: a computational and analytical node, a processing control system, a data storage system, and a data processing and analysis platform. The computational node is a dedicated node that has a set of a specific software to process and analyze certain data sets. This node has an API-enabled connection to the Internet.

Microservices can be located anywhere on the Internet and have different software and hardware characteristics.

The processing control system serves to organize a single interface to access heterogeneous processing services, as well as for managing the process and access sharing to various services.

The data storage system is responsible for access to processing and analysis results, as well as to the data uploaded by users for further processing.

The processing platform provides a set of interfaces for interaction with the processing system, including a web interface, and provides an application programming interface for interaction in the system-system schem

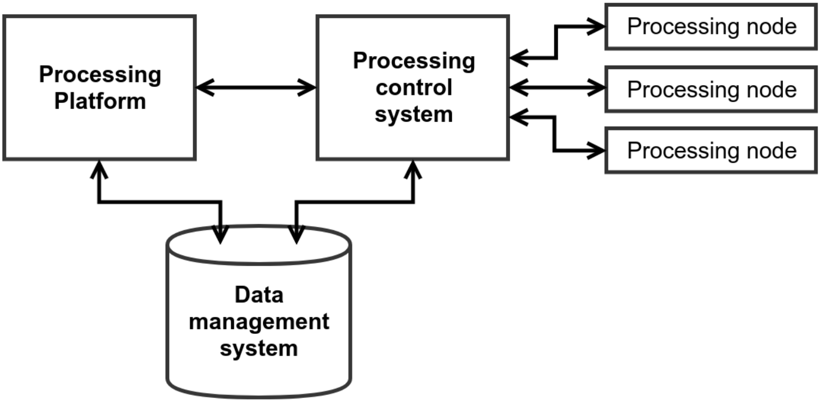


Figure 1. The general scheme of the computational-analytical environment.

Functional Links Between Objects

The general scheme of the computational-analytical environment is shown in Figure 1.

Environment Implementation Main Technologies

At present, when creating such systems, the approach based on the technology of web-services is actively applied. This technology allows to access the software system using a standardized interface via the Internet. General standard for initiation and management of processing procedures in the framework of web services technology is the OGC Web Processing Service (OGC WPS) interface (OGC Web Processing Service, http://www.opengeospatial.org/standards/wps) developed by the international organization for standardization of the Open Geospatial Consortium. This interface is implemented on the basis of the HTTP protocol (HTTPS), is widely used in various scientific data processing systems [*Avramenko and Fedorov,* 2014; *Bychkov et al.,* 2014; *Fedorov and Shumilov,* 2015]; a variety of software products support the operation with geoinformational services.

Processing Environment Architecture

Based on the stated requirements, the architecture of the distributed geological data processing environment was developed (Figure 2), as well as the functional requirements for each component of the environment were formulated.

Computing-Analytical Nodes

Since we proposed to use WPS-processes as an interface to access external services, a number of requirements are imposed on the nodes:

- Permanent IP address (or domain name) and a port

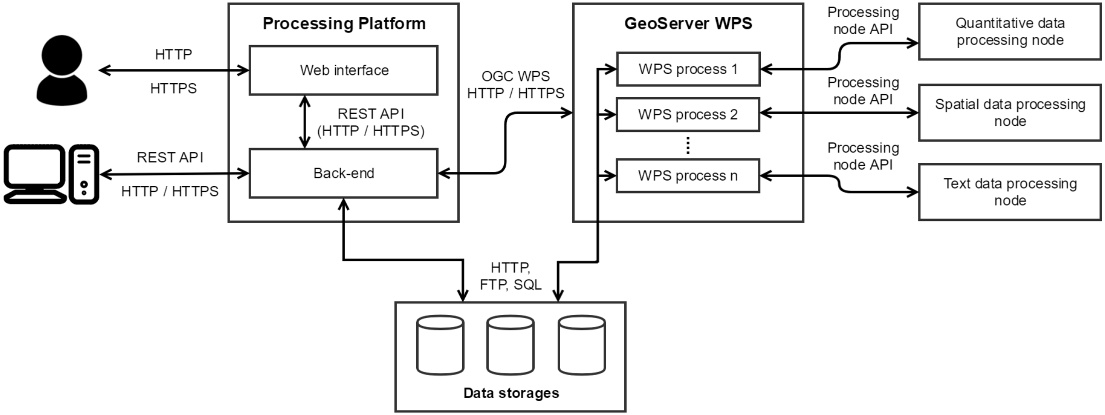


Figure 2. The architecture of the distributed geological data processing environment.

with external access via the Internet;

- Client program interface (API) and libraries implementing it in the Java language, or via the HTTP / HTTPS protocol (REST or SOAP);
- Ability to start the processing procedure with the specified parameters;
- Ability to obtain the result of processing in text or binary formats, including in the form of URL links;
- Ability to work with user data in one of the following ways:
  - Reading remotely placed data by URL address
  - Temporary loading data to the computing node
  - Transferring data in binary format as a processing parameter

Processing Control System

The main functions of the processing management system are: management of the processing (launch, status tracking, return of the processing result), delineation of access to various services and control over the use of resources. Open source software package GeoServer was selected to create and host your own WPS processes.

Thus, for each external service, a separate WPS process is created with the corresponding startup parameters (Figure 3).

Using WPS as an access interface to remote services allows you to execute several computational processes sequentially. Thus, outcomes of one or several processes can be used as input parameters for another process, thereby providing the possibility to combine the results of processing different types of data.

Processing Platform

The processing platform is the interface between the user and the processing control system. It is a client-server software solution that provides access to external systems using the REST interface, as well as a web interface for accessing the platform through a web browser.

The main features of the processing platform are:

- Uploading user data to the repository;
- Providing a list of services that support user-selected data for processing;
- Providing the ability to manage the processing parameters for the selected service;

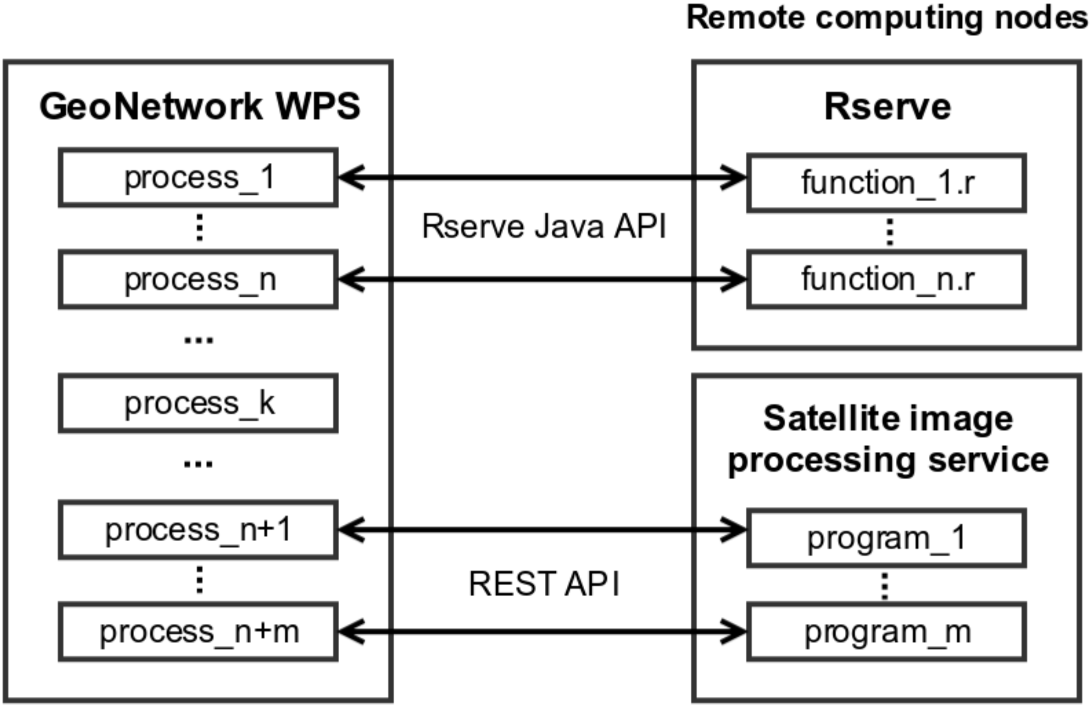


Figure 3. Interaction scheme between WPS processes and external computing nodes by the example of and the satellite data processing system.

- Ability to start processing the selected service with the specified parameters;
- Getting the result of processing as a URL link.

Processing Chains

Often, to address certain goals, it is necessary to execute sequentially several processes, where the result of one process is transferred as an input parameter for another process. Using the functionality of WPS, you can create a sequence of processes, called processing chains, which are a combination of existing services, defining the relationships and interactions between them, aimed at solving a specific problem.

According to the functionality of the processing chain we can distinguish auxiliary, systematic and thematic chains.

Thematic chains are designed to solve problems requiring the use of several processing and analysis methods. As an example, we can include the use of spectral remote sensing techniques (multispectral and hyperspectral surveying) to determine the composition of geological formations; to monitor the state and modification of anthropogenic formations, etc.

Auxiliary chains – perform auxiliary tasks, such as format conversion, scale transformation, primary data visualization, etc.

Systematic ones – perform the task of monitoring the computing nodes, the availability of services, their description correctness, etc.

Development of a Software Prototype of a Distributed Computational and Analytical Environment for the Scientific Study of Geological Systems

The platform acts as an intermediary between the user and external processing systems and provides a single access interface to all processing algorithms available in external processing systems (system nodes). Furthermore, the described architecture assumes the possibility of using data not only from open sources available in the system, but also loading processing data by the user. In the developed environment, the computational and analytical blocks for processing and analyzing geological information are organized as sets of service and analytical functions with an ability of a processing methods choice. Processing methods include data loading, transformation of formats, methods of analysis and visualization of results.

Access to the processing and analysis units is provided through the distributed data services management platform. As an intermediate interface between the system and its external nodes, the OGC WPS interface is used.

Computing Nodes Designed for This Environment:

Processing Node of Quantitative Data Charts.

When processing geological quantitative data charts, methods of statistical and multidimensional data analysis are required. The set of possibilities of the computing environment R is sufficient to select it as a computational node software.

In constructing this node, besides the main functions of statistical data processing and construction of elementary charts we use the following methods: cluster analysis, principal components, factor analysis, discriminant analysis, canonical variables analysis, linear and nonlinear regression analysis, multidimensional scaling [*Platonov and Naumova,* 2017]. The data processing block has a scalable infrastructure to perform the computation of the analysis of the geological quantitative datasets (Figure 4). The authors have selected the R programming language to enable this scalability. This language is famous for its flexibility to add proprietary algorithms.

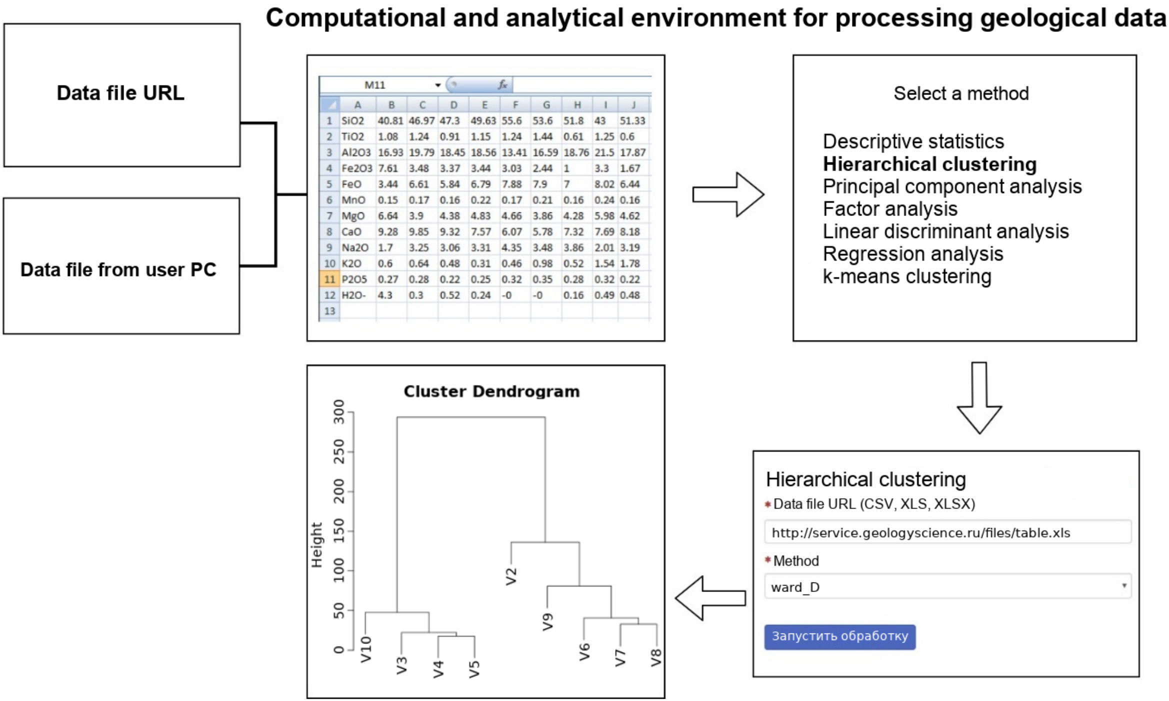
The “Rserve” extension allows other programs to use the capabilities of the R language via the TCP/IP protocol. Each connection has a separate workspace and a directory for loading data. The computing is accessible by IP address (or domain name).

Spatial Data Processing and Analysis Node.

On this node, the following operations are possible for vector data in GIS-formats: definition of the relative location of objects, getting inside objects, finding the nearest objects, intersection; calculation of area, length, perimeter, distance between objects etc.

Satellite Data Processing Node.

As part of the work on the implementation of this computing node, satellite data processing services are being developed: projection transformations, atmospheric correction, consolidation of satellite imagery, conversion of satellite data into PNG, GeoTIFF, ASCII Grid,



Computational and analytical environment for processing geological data

Data file URL

Data file from user PC

	A	B	C	D	E	F	G	H	I	J
1	SiO2	40.81	46.97	47.3	49.63	55.6	53.6	51.8	43	51.33
2	TiO2	1.08	1.24	0.91	1.15	1.24	1.44	0.61	1.25	0.6
3	Al2O3	16.93	19.79	18.45	18.56	13.41	16.59	18.76	21.5	17.87
4	Fe2O3	7.61	3.48	3.37	3.44	3.03	2.44	1	3.3	1.67
5	FeO	3.44	6.61	5.84	6.79	7.88	7.9	7	8.02	6.44
6	MnO	0.15	0.17	0.16	0.22	0.17	0.21	0.16	0.24	0.16
7	MgO	6.64	3.9	4.38	4.83	4.66	3.86	4.28	5.98	4.62
8	CaO	9.28	9.85	9.32	7.57	6.07	5.78	7.32	7.69	8.18
9	Na2O	1.7	3.25	3.06	3.31	4.35	3.48	3.86	2.01	3.19
10	K2O	0.6	0.64	0.48	0.31	0.46	0.98	0.52	1.54	1.78
11	P2O5	0.27	0.28	0.22	0.25	0.32	0.35	0.28	0.32	0.22
12	H2O-	4.3	0.3	0.52	0.24	-0	-0	0.16	0.49	0.48
13										

Select a method

Descriptive statistics
Hierarchical clustering
Principal component analysis
Factor analysis
Linear discriminant analysis
Regression analysis
k-means clustering

Hierarchical clustering

* Data file URL (CSV, XLS, XLSX)

http://service.geologyscience.ru/files/table.xls

* Method

ward_D

Figure 4. Script of processing the chart of geological quantitative data by the method of hierarchical cluster analysis.

PRO; streaming data processing (median smoothing, data profile extraction, filtering and data retrieval), etc.

Visualization Node.

At this node, users will be given the opportunity to build various types of graphs and raster maps.

When interacting with this “3D Earth” service, the user will be able to see on the Earth’s surface the geographic location of geological objects, sections, and so on. It will also be possible to combine the analyzed material with geological maps, museum artifacts, etc., as well as with global movements of lithospheric plates at different time intervals (including the ability to view their movements between selected time intervals in millions of years).

External Computing Nodes.

An important component of the Environment is geographically distributed external computational and analytical nodes. Nowadays there are dozens of services that solve various problems. If the service meets the requirements set for the computational and analytical nodes, it can be integrated into the Environment as an external processing and analysis node. In particular, access to the Cermine Web service [*Tkaczyk,* 2015] is realized. This service extracts metadata from scientific publications. In future, it can be integrated with ESRI cartographic services, USGS geological services, Rosnedra agency, services of Geophysical Center of the Russian Academy of Sciences, Institute of Volcanology and Seismology of FEB RAS, Pacific Institute of Geography FEB RAS, etc.

Interaction With Users

The user does not know the internal logic of the computing-analytical node, he/she only focuses on its description and interacts with it only through its standardized interface.

Computational and analytical capabilities of the environment can be used by users or external information systems based on their own information when accessing the platform for managing computer services.

Distributed computing-analytical environment is developed within the framework of the Information and Analytical Environment for scientific study of geological systems [*Naumova et al.,* 2017]. At present, the computational and analytical environment is under implementation. In the Environment, the following services operate on a trial basis: regression analysis, principal component analysis, hierarchical cluster analysis, k-means, factor analysis, linear discriminant analysis, canonical variables analysis and construction services for various charts, including triple diagrams widely used in geology. Integration with a number of remote services is implemented, such as automatic extraction of metadata and various information blocks from the publications, and etc.

References

Avramenko, Yu. V., R. K. Fedorov (2014), WPS-services for processing remote sensing data, *Bulletin of Buryat State University*, no. 9-1, p. 12–15.

Bychkov, I. V., G. M. Ruzhnikov, R. K. Fedorov, A. S. Shumilov (2014), Components of WPS-services for geodata processing, *Bulletin of Novosibirsk State University. Series: Information technology*, *12*, no. 3, p. 16–24.

Fedotov, A. M., V. B. Barakhnin, A. E. Guskov, Yu. I. Molorodov (2006), Distributed information and analytical environment for environmental systems research, *Computational Technologies*, *11*, no. Special, p. 113–125.

Fedorov, R. K., A. S. Shumilov (2015), Creation and publication of WPS-services on the basis of cloud infrastructure, *Bulletin of*

BSU, no. 4, p. 29–35.

Naumova, V. V., S. V. Dyakov, K. V. Platonov, V. S. Eremenko, B. V. Shuvalov (2017), Information and Functional Capabilities of the Internet in the Far East, *International Conference "Mineral prospectivity: current approaches and future developments in predictive geosciences", 24–26 October 2017, BRGM, Orléans (France). Book of Abstracts*, p. 22–24, BRGM, Orléans, France.

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

Tkaczyk, D., P. Szostek, M. Fedoryszak, P. Dendek, L. Bolikowski (2015), "CERMINE": automatic extraction of structured metadata from academic literature, *International Journal on Document Analysis and Recognition*, p. 317–335, **Crossref**

Tolpin, V. A., et al. (2011), Creation of interfaces for working with data of modern remote monitoring systems (GEOSMIS system), *Modern Problems of Earth Remote Sensing From Space*, *8*, no. 3, p. 93–108.