Methods and Technologies for Integration and Processing of Geographically Distributed Quantitative Geological Information

© K.A. Platonov
Vernadsky State Geological Museum,
Moscow, Russia
twinco@mail.ru

Abstract. There are suggested methods and technologies for integration and processing of geographically distributed quantitative geological information. The methods based on DataCite approaches, DOI-system, OAI standards and protocols are applied. The developed and adapted methods and technologies were taken as principles in creation of Information system for integration and processing of quantitative geological information.

Keywords: integration of quantitative data, standardization of quantitative data, processing node.

1 Introduction

A large geological empirical material is generated and published annually on the Internet - quantitative determinations of chemical and mineral composition of rocks, ores, minerals and their aggregates in the Russian Federation. The data are placed at the spot of their receipt and the most intensive use in databases and information systems, world networks for the exchange of scientific data, as well as in scientific journals and monographs. Geographically distributed information resources make it difficult for geologists to obtain complete, reliable and sufficient information for solving the scientific and production tasks assigned to them.

At the present, the organization of a single point of access to geographically distributed quantitative information through a single unified interface is an urgent problem for organization of information support and support of scientific geological research in the Russian Federation. For a long time the integration of quantitative information in geology was carried out at the physical level, in particular the consolidation of data with subject and territorial constraints.

In 2014 the final "Declaration of Data Citation Principles" and the mechanism for data publishing as an independent unique product of scientific work were published [3].

In 2010-2014 the developers of the DataCite project proposed a software implementation based on the "Declaration of Data Citation Principles". The key point of the project is the creation of a DOI registration agency for scientific data [1]. The procedure for quantitative datasets publishing includes a DOI identifier assignment, a unified meta description creation and registration in the central DataCite repository (http://www.datacite.org/). Datasets are placed in specialized systems supporting OAI metadata exchange protocols and sending the current information to the central DataCite repository.

The appearence of permanent data identification system and a uniform global metadata scheme allows to organize the integration of information contained in geographically distributed sources at a logical level.

The basic principles of information systems(IS) of integration and data management organization are described in the work [16]. Each IS should support maintain conditions for being Findable, Accessible, Interoperable, and Reusable data. The uniqueness of the stored digital objects is based on the "Declaration of Data Citation Principles".

The main sources of tables of geological quantitative data are listed in the table(see Table 1).

Table 1 Sources of quantitative geological information

Sources	Access protocols	Data adding form	Metadata format	Storage form
Repositories	HTTP,OAI	Manually oai-pmh 2.0	oai_dc	Full texts of publications
Database "GeoRoc" [12]	HTTP	Manually	N/A	Table files
FAIR Systems "Pangaea"[5]	API,HTTP,OAI	Manually	oai_dc, pan_md datacite3	Metadata and datasets
DataCite	API,HTTP,OAI	oai-pmh 2.0	oai_dc, oai_datacite	Metadata

---

Proceedings of the XX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2018), Moscow, Russia, October 9-12, 2018

2 Methods and technologies for the integration and processing of quantitative information in geology

In the frames of the work on development of the Information infrastructure of support and follow up of the scientific geological research[8] , was targeted the goal to develop methods and technologies for creation an information system for the integration and processing of geographically distributed quantitative geological information.

The following tasks were formulated:

- Development of methods and technologies for integration quantitative data from various technological, geographically distributed sources: repositories, scientific journals and monographs, information systems and databases, scientific information exchange networks;
- Development of methods and technologies for generation metadata of the received tables of quantitative data in international formats;
- Implementation of a system for integration and processing geographically distributed quantitative geological information;
- Creation of thematic processing block of geological information.

2.1 The integrated method of quantitative information tables extraction from scientific publications

Scientific publications have long been perceived as "Big Data." Therefore, the tasks of this information flow processing are solved with a high level of automation and availability of horizontal scaling of the used algorithms.

The structure of the scientific publication is analyzed in [15]. The process of automatic extraction of textual information from files in PDF format is described in the same work: article's metadata and list of references. The present work became a prerequisite for proposed by the author integrated method of quantitative data tables extraction from scientific publications.

The PDF format, unlike other formats, does not contain data about any structures in the document. The file stores information about the output to the right places of symbols, lines, curves, rectangles, raster images and other geometric primitives. Thus, the task of finding a table in PDF format on an article page is similar to the tasks of recognition a table structure on a raster image. The technology for obtaining quantitative data in the form of a publishing electronic table consists of 5 stages: page layout analysis, table structure detection, structure recovery, cell function analysis, header acquisition and table notes.

The following methods were considered for the problem solving:

- RLSA - the essence of the method is to create a binary mask image in the vertical and horizontal direction. Further, the masks are combined and, based on the frequency characteristics of the black and white pixels, the page layout is restored. [17];
- X-Y Cut - the essence of the method is that the page is alternately divided into blocks by a horizontal or vertical cut. The result is a tree-like structure of the page, where the whole page is the root, and the related blocks of the layout are the descendants. [9]
- Segmentation using the maximum white rectangles - the method searches for all high maximum white rectangles and evaluates as candidates in between the column separators. In accordance with the found column structure, strings, paragraphs and other elements of the page are searched [2].
- Docstrum - a method designed for segmentation of textual information. All characters are clustered by size (the author suggests 2 clusters). The distance from the symbol to 4-5 of its neighbors is calculated. A histogram of distances is constructed. The first three peaks mean: the distance between the letters in the word, the line spacing, the space. We combine symbols in words, words in strings, lines in paragraphs. The result is a layout [10].
- The Voronoi diagram. The page is divided into areas, each of which corresponds to one reference point (the center of the symbol) and is the set of points of the plane for which the given reference point is closer than any other reference point [7].

It should be noted that all algorithms are designed for text search for the subsequent application of algorithms for optical character recognition. In the case of a PDF file, this is not required. Using the PDFminer software [11], a digital version of the characters and coordinates of the area they occupy on the page are available. The transformation of the PDF-format into a bitmap is done using a virtual printer.

I proposed an integrated method for extraction quantitative information tables from the publication text in PDF format, which performs a complete cycle from segmentation the page to receipt an electronic version using a series of related methods.

For each stage, the adapted versions of the above algorithms were used. The analysis of the page layout and table detection is carried out by a combination of RLSA and X-Y Cut algorithms. The difference was the use of 3 gradations (white, gray, black) when winding the lines in a horizontal and vertical direction. A modified version of the X-Y Cut method for each column builds a block tree (text, table, image, graph, etc.).

At the stage of restoration the structure of the table, the segmentation method is used using the maximum white rectangles. In my case, the maximum white and black rectangles are searched, depending on the presence of internal boundaries.

Finally, using the Docstrum method, the values of cells are extracted and their function is analyzed (a cell with a value, the name of a column or rows, etc.)

Regular expressions and keywords are used to find the table header and notes: Table and Note.

The test sample consisted of 136 scientific publications. 259 objects are extracted. The error percentage was 5%. Part of the tables cannot be presented in a spreadsheet format due to problems with the encoding of Russian characters in PDF files.

Integration of quantitative information from FAIRness sources and data networks is carried out in the standard mode of metadata exchange and OAI protocol data.

2.2 Storage of geological tables of quantitative data

In the Earth sciences, in particular in geology, quantitative tables of experimental data are accompanied by geological, geographic, temporal and analytical descriptions. This information can be contained in the header of the table, a note, and in other elements of the table, as well as in its metadata. For optimal organization of storage and development of a topical search, the author proposes to catalog the tables of quantitative geological data using this information. The extracted textual geographic characteristics are important for the subsequent obtaining of the absolute coordinates of the geological object. The author proposes to extract geological descriptions using three developed thesauri: names of geological complexes, geologic time scale, names of mineral deposits. A textual description of the analytical method of geological data analyzing (data on the instrument, the method of analysis, the location and time of the study) is in the note and is easily extracted.

Storage of tables, retrieved from publications, is carried out in XML-form which repeats the structure of the table, the title and the notes to the local storage system organized as a relational database. Each quantitative data set is then assigned a unique identifier (DOI) and its card is generated according to the Dublin Core metadata format specification or its modified version used in the DataCite project. In 2017, the DataCite working group published a 4.1 version of the metadata schema for publishing and citing research data [4]. According to the documentation, the metadata should contain three levels of properties: mandatory, recommended and optional. There are enough mandatory properties to include records in the DataCite database, which include information about identifiers, collection header, data about the author, publisher and the year of publication, type or format of the resource. To improve search and integration properties of the data sets it is necessary to fill in the recommended additional properties, such as subject area, keyword, members, description, additional identifiers, geolocation (point, rectangle or polygon), the publication or project data.

The author developed an algorithm that compares the information from the metadata of the publication (author, a link to the publication), geological (subject area), geographic (geolocation), temporary and analytical descriptions of the table to generate metadata automatically.

The author proposes to store the metadata obtained this way in the IS metadata base which will be available to external ISs via API and OAI protocols.

2.3 The system of integration and processing of geographically distributed quantitative geological information

The system for integration and processing geographically distributed quantitative geological information was developed and implemented basing on the proposed methods. Unlike the previous implementation of the system [12-13], in the new version there are two blocks: the integration unit and the geological information processing unit. The first - performs the integration of quantitative data, their primary processing, storage, retrieval, external exchange with other systems on standard protocols and information management. The second block is computationally analytical, processing and analyzing quantitative tables of geological data. This separation is required to solve the problems of scaling the system's capabilities while processing and analyzing scientific geological data, including BigData.

The following tasks are solved:

- Organization of the mechanism for the integration of quantitative data and their descriptions from different types of geographically distributed sources: scientific publications, world databases, data networks;
- Development of a system for storage geological quantitative data and metadata;
- Creation of an algorithm for generation metadata in the DataCite format;
- Development of algorithms for automatic extraction of quantitative data from sources.
- Ensuring the availability of data in user-specified formats through the search system and catalogs
- Ensure availability of data through the API and the Open Archives Initiative (OAI) protocols.
- Development of metadata management services: data cataloging, packet loading, synchronization, monitoring and usage statistics, etc.
- Organization of a block for the subject processing of geological quantitative information.

The general functional scheme of the System is shown in figure (see Figure 1).

The system provides data input in two modes: automatic and manual.

The user through the Search Module performs a query to the System in one of three modes: simple, extended and spatial. The provisioning module generates a response and makes it possible to obtain a card of the quantitative data table with all output information in PDF and Excel formats (see Figure 2).

The representation module manages the catalog system, which is organized according to a thematic principle. The field catalog contains more than 10 000 titles (the list of Cadastre Deposits of the Russian Federal Geologic Fund "ROSGEOLFOND"). The catalog of geological complexes of the Russian Far East includes 540 items.

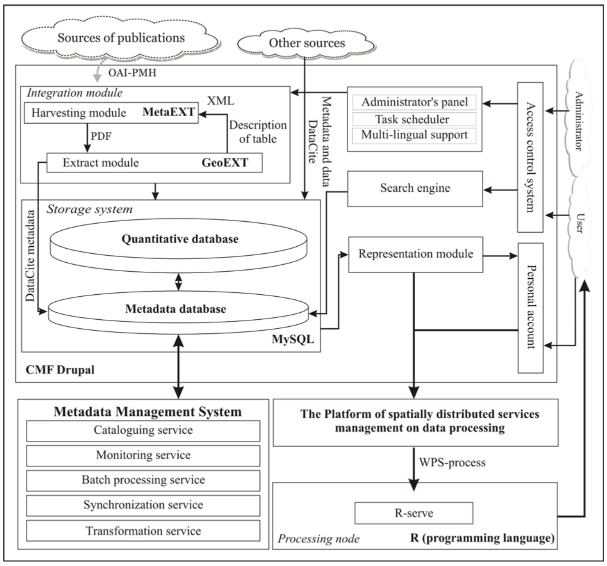


Figure 1 Functional scheme of System.

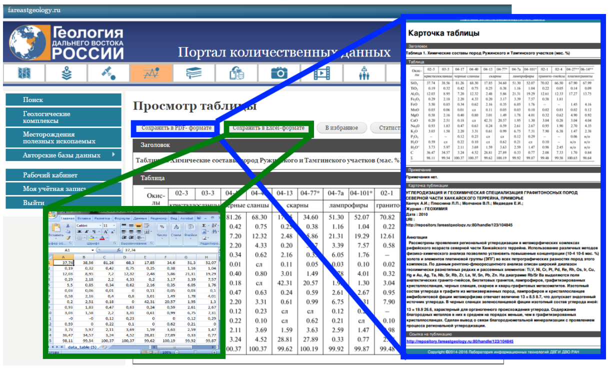


Figure 2 An example of data provision for users in PDF format (on the right) and in Excel format (on the left)

The user can access to the Personal account, which stores the marked sets of quantitative data and the results of their processing. Through the Personal account is a manual entry of information. Through the administrator's unit, all processes and services of the System are managed and configured; the frequency and time of services start-up, the status and use of the resource are monitored.

2.4 Block for processing and analyzing of geological quantitative information

While carrying out geological studies, the need to apply quantitative methods of processing, analysis and generalization of data increases. The complexity of the methods used is increasing from the methods of elementary statistics to the multidimensional analysis of data and the construction of models of natural processes.

Therefore, the actual task is to provide the researcher with tools for processing and analyzing quantitative data on the Internet. Moreover, the user-geologist is interested in the possibility of processing and analyzing personal data by various methods not on his own PC, but on the side of remote servers and, if necessary, using supercomputers.

In the system under development, a computational and analytical unit for processing and analyzing geological quantitative information is organized in the form of a set of service and analytical functions with the possibility of user access to the selection of the processing method; chain processing, including data loading, transformation formats, method analysis and visualization of results; a thematic chain that includes a sequence of analysis methods.

While processing geological tables of quantitative data, methods of statistical and multidimensional data analysis, as well as the construction of graphs and maps, are in demand. The set of possibilities of the computing environment R is sufficient to select it as a computational node software. In addition to the basic functions of statistical data processing and the construction of elementary graphs, we use the following methods: cluster analysis, principal components, factor analysis, discriminant analysis, canonical variables analysis, linear and nonlinear regression analysis, multidimensional scaling, as well as specialized sample visualization packages, for example, "GEOmap", developed by Jonathan Lees [6] and others. The computational capabilities of a node of a computational-analytical block for processing and analyzing quantitative geological data are extensible, since the medium R chosen by the author offers a flexible system for adding author's 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 node is accessible by IP address (or domain name).

The block is accessed through the distributed data services management platform. The platform uses the Web Processing Service (WPS) standard created by the Open Geospatial Consortium to exchange information between spatial data services, but the WPS protocol is universal and can be used to unify access to compute nodes with any type of data. (see Figure 3).

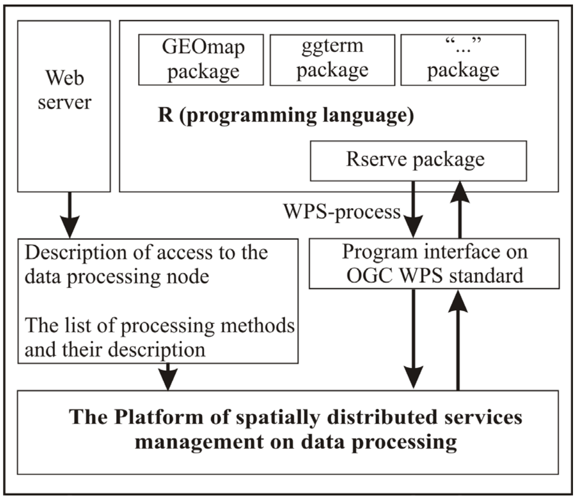

Web server
GEOmap package
ggterm package
"..." package
R (programming language)
Rserve package
WPS-process
Description of access to the data processing node
The list of processing methods and their description
Program interface on OGC WPS standard
The Platform of spatially distributed services management on data processing

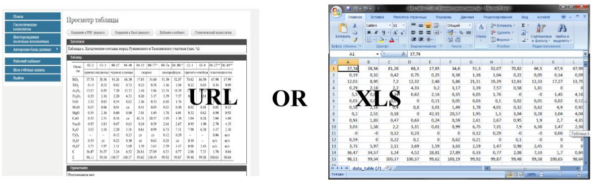
Figure 3 Schema of the data processing node in R language

The platform provides a single access interface to all registered processing algorithms and computing resources and acts as an intermediary between the user and external processing systems.

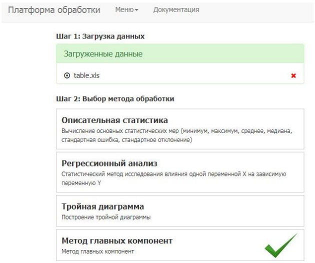
A script for one’s access to the analysis methods (principal component analysis) is shown in figure (see Figure 4). The user downloads the data (see Figure 4(a)) and selects the analysis method (see Figure 4(b)). The management platform via the WPS process calls remote processing on the compute node and returns the result in Excel format (see Figure 4(c)).

The Computational and Analytical Node of the System can be used by users or external ISs based on their own information when accessing the Platform for Management of Computing Services.

Currently, the system is in the process of finalization and testing. After the completion of these works it will be available for the Internet users.



(a) Select data



Шаг 1: Загрузка данных

Загруженные данные

table.xls

Шаг 2: Выбор метода обработки

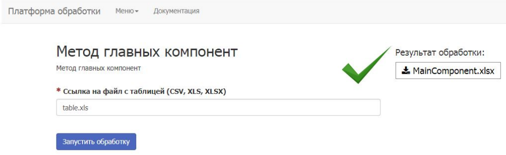
Описательная статистика
Вычисление основных статистических мер (минимум, максимум, среднее, медиана, стандартная ошибка, стандартное отклонение)

Регрессионный анализ
Статистический метод исследования влияния одной переменной X на зависимую переменную Y

Тройная диаграмма
Построение тройной диаграммы

Метод главных компонент ✓
Метод главных компонент

(b) Select principal component analysis



Метод главных компонент

Метод главных компонент

* Ссылка на файл с таблицей (CSV, XLS, XLSX): table.xls

Результат обработки: MainComponent.xlsx

(c) Start processing and get the result

Figure 4 Script for processing the quantitative table by the principal component analysis

3 Conclusion

Methods and technologies for integrating tables of quantitative data from different technological, geographically distributed sources are proposed.

An integrated method for extracting tables of quantitative data from the texts of geological scientific publications is developed.

An algorithm for automatically generating metadata in the DataCite format is developed.

Basing on the developed and adapted methods and technologies, a system for the integration and processing of geographically distributed quantitative geological information is implemented.

A computational and analytical block for processing and analyzing geological quantitative information is developed and implemented.

Acknowledgments. This work is supervised by Professor(Geology and Mineralogy) Naumova Vera Viktorovna, Head of department, Vernadsky State Geological Museum.
The work is supported by RFBR, №14-07-00068.

References

[1] Brase, J.: DataCite - A global registration agency for research data. In: Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology, pp. 257-261. IEEE, Beijing (2009). doi: 10.1109/COINFO.2009.66

[2] Breuel, T.M.: Two Geometric Algorithms for Layout Analysis. In: Lopresti, D., Hu, J., Kashi, R. (eds) Document Analysis Systems V. DAS 2002. Lecture Notes in Computer Science, vol 2423. Springer, Berlin, Heidelberg (2002), doi: 10.1007/3-540-45869-7_23

[3] Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone, M. (ed.) San Diego CA: FORCE11; (2014) doi:10.25490/a97f-egyk

[4] DataCite Metadata Working Group.: DataCite Metadata Schema for the Publication and Citation of Research Data. Version 4.1. DataCite e.V. (2017) doi: 10.5438/0015

[5] Diepenbroek, M., Grobe, H., Reinke, M., Schindler, U., Schlitzer, R., Sieger, R., Wefer, G.: PANGAEA - an information system for environmental sciences. In: Computers & Geosciences, vol. 28 no. 10,pp. 1201–1210. (2002) doi:10.1016/S0098-3004(02)00039-0

[6] GEOmap package | R Documentation. https://www.rdocumentation.org/packages/GEOmap/versions/2.4-4

[7] Kise, K., Sato, A., and Iwata, M.: Segmentation of page images using the area voronoi diagram. In: Computer Vision and Image Understanding, vol. 70, issue 3, pp. 370-382. (1998) doi:10.1006/cviu.1998.0684

[8] Naumova, V. V., Goryachev, I. N., Dyakov, S. V., Belousov, A. V., Platonov, K. A.: Modern technologies of development of the Information infrastructure to support the research on geology of the Russian Far East. In: Information Technology, vol. 21, no. 7,pp. 551–559. (2015) (in Russian)

[9] Nagy, G., Seth, S., and Viswanathan, M.: A prototype document image analysis system for technical journals.In: Computer, vol. 25, no. 7, pp. 10-22. (1992). doi: 10.1109/2.144436

[10] O'Gorman, L.: The document spectrum for page layout analysis. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15, no. 11, pp. 1162-1173, (1993). doi: 10.1109/34.244677

[11] PDFminer. https://www.unixuser.org/~euske/python/pdfminer

[12] Platonov, K. A.: Methods and technologies for creation of the information processing system applied to publications on geology of the Russian Far East. In: RUSSIAN JOURNAL OF EARTH SCIENCES, VOL. 15, ES4005,(2015). doi:10.2205/2015ES000560

[13] Platonov, K.A., Naumova, V.V.: Methods and technologies for geological quantitative information integration. In: Proceedings of Irkutsk State Technical University, vol. 21, no 21, pp. 67–74, (2017). (In Russian) doi: 10.21285/1814- 3520-2017-2-67-74

[14] Sarbas, B.: The GEOROC Database as Part of a Growing Geoinformatics Network. In: Geoinformatics 2008 - Data to Knowledge, Proceedings, pp. 42-43. Potsdam, (2008)

[15] Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P. J., Bolikowski, Ł.: "CERMINE": automatic extraction of structured metadata from scientific literature. In: International Journal on Document Analysis and Recognition, pp. 317-335. (2015). doi: 10.1007/s10032-015-0249-8

[16] Wilkinson, M. D. et al.: The FAIR Guiding Principles for scientific data management and stewardship. In: Sci. Data 3:160018 (2016). doi: 10.1038/sdata.2016.18.

[17] Wong, K. Y., Casey, R. G., and Wahl, F. M.: Document Analysis System. In: j-IBM-JRD, vol.26, no. 6, pp. 647–656. (1982).