RUSSIAN JOURNAL OF EARTH SCIENCES
vol. 15, 4, 2015, doi: 10.2205/2015ES000560

Methods and technologies for creation of the information processing system applied to publications on geology of the Russian Far East

K. A. Platonov

Far East Geological Institute FEBRAS, Vladivostok, Russia

Abstract. Quantitative database extracted from scientific publications is developed within the framework of the development of the Information Infrastructure in order to support scientific geological research of the Russian Far East. Based on the analysis of the existing methods and technologies used for processing of scientific publications, new technological approaches to automatic extraction and processing of quantitative data in scientific publications are applied.

This is the e-book version of the article, published in Russian Journal of Earth Sciences (doi:10.2205/2015ES000560). It is generated from the original source file using LaTeX's **epub.cls** class.

Introduction

Lately, a close attention has been drawn upon Information systems that operate on scientific publications. The total amount of information stored in such systems is more than millions of articles. The concept of "open excess" to scientific data institutionally adopted by international community in 2004 has facilitated user access to full-text publications. Web sites of journals and publishers, full-text databases, electronic catalogs of academic libraries, digital repositories, e-libraries and academic networks are considered as a new type of information sources, i.e. scientific publications. As a result, Information science has to face challenges of big data processing in order to obtain new results in different scientific fields.

Structure of a Scientific Publication

A scientific paper is a written and published report describing original research results [Day, 1979]. The content of a publication can be divided into several zones [$Tkaczyk\ et\ al.$, 2015]: metadata, body and references (Figure 1). Metadata information includes such simple metadata types as title, abstract, key words, personal

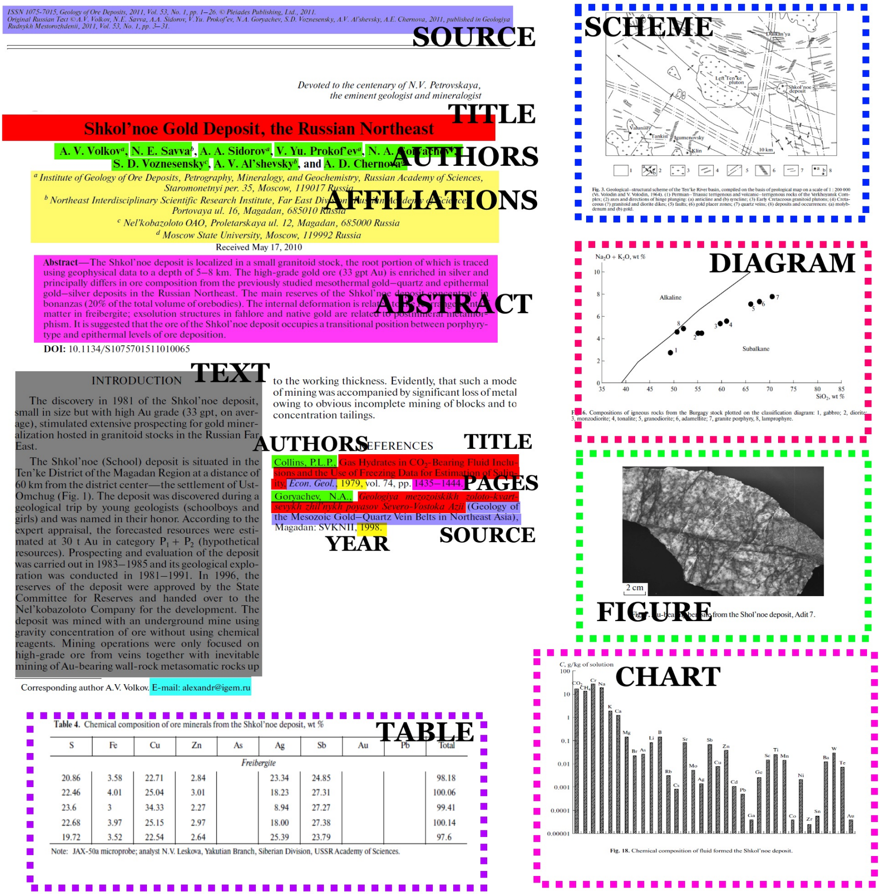


Figure 1. Main zones of a scientific publication.

records of the author, such as names and affiliation, file format, unique identifiers, place of publication and other information. Metadata information is located on the cover page of the article. Zone of references contains citations to articles, books, e-resources or web pages, public or legal documents etc.

Body zone contains a combination of different data types such as publication's text, images, charts, graphs, tables, diagrams etc. Thus, the objective is to localize, extract, store and process the relevant data. Creating databases by extracting similar objects from scientific publications makes it possible to apply modern methods and approaches of information processing to aggregated samples of these objects and thereby achieve new results.

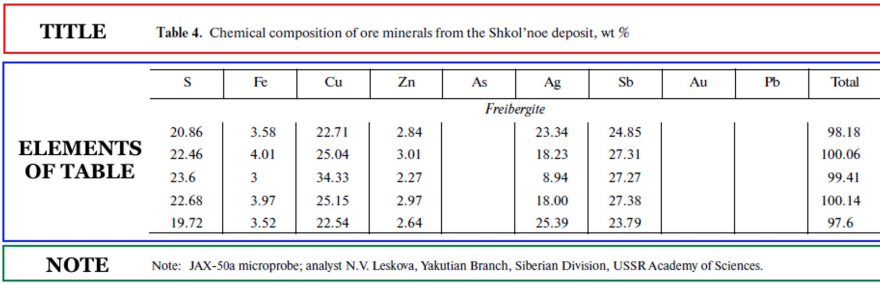
Storage Systems for Scientific Publications

Digital repositories, academic e-libraries and full-text databases have turned into specialized storage systems for large amounts of scientific publications. These systems keep to the world standards of data description, provide common (user) access and allow to work in the

automatic mode.

At present the number of repositories and digital libraries as well as the number of publications on geosciences is growing rapidly. According to the information provided by digital e-library eLibrary.ru more than half a million of publication on geology written by Russian authors is registered in its database, 32,000 of these publications contain one or more tables. Most tables inserted in geological articles displays quantitative data collected by means of analytical study of samples.

Systems based on extraction of quantitative data from scientific publication are being developed actively all over the world. "GEOROC" system (http://georoc.mpch-mainz.gwdg.de/georoc/) developed under the supervision of Baerbel Sarbas in the Max Planck Institute for Chemistry contains more than 422,000 analyzed samples retrieved from 13,900 publications. [$Sarbas$, 2008]. Internet resource "American Mineralogist Crystal Structure Database" (http://rruff.geo.arizona.edu/AMS/) contains quantitative data published in four journals: American Mineralogist, The Canadian Mineralogist, European Journal of Mineralogy, Physics and Chemistry of Minerals. The project is developed by the Department of Earth Sciences, University of Arizona [$Downs\ and$



S	Fe	Cu	Zn	As	Ag	Sb	Au	Pb	Total
					Freibergite				
20.86	3.58	22.71	2.84		23.34	24.85			98.18
22.46	4.01	25.04	3.01		18.23	27.31			100.06
23.6	3	34.33	2.27		8.94	27.27			99.41
22.68	3.97	25.15	2.97		18.00	27.38			100.14
19.72	3.52	22.54	2.64		25.39	23.79			97.6

Figure 2. 3-zone fragmentation of quantitative data table (top to bottom): title, elements of table, note.

Hall-Wallace, 2003].

The use of standardized storage systems for scientific publications enables to perform automatic data extraction, in particular table data from the body of the publication with the help of modern methods.

Methods of Processing Scientific Publications

There are two main approaches to detecting and extracting table data from born-digital documents. If publications or documents are bitmap images, the first step is to use methods focused on automatic detection of table boundaries, structure and cells, the second step is to use Optical Character Recognition (OCR) [*Bart*,

2012; *Seo et al.*, 2015]. The disadvantage of this approach is that the entire table detection process may show bad results if there are no cell separators or if it is impossible to use specific properties of table cells. Poor performance of the OCR technology also makes it difficult to ensure accurate data extraction.

Table reconstruction method is based on the usage of text blocks [*Hassan and Baumgartner*, 2007; *Shigarov*, 2009; *Tkaczyk et al.*, 2015]. In this approach, individual characters are merged into non-overlapping blocks, blocks of white spaces and more complex structures: cells, rows, columns, table regions etc. Block order reconstruction is based on the rules of cell layout and table structure (margin, gaps and intervals between table elements, box drawing characters and separator lines etc.). Complexity of the table structure influences the accuracy of data extraction.

Creation of the Data Portal on Geology of the Russian Far East

Within the framework of the development of the Information infrastructure and in order to support scientific geological research of the Russian Far East [*Naumova*

et al., 2015], we have to deal with the necessity to create quantitative database.

The digital repository "Geology of the Russian Far East" (http://repository.fareastgeology.ru) [*Naumova and Belousov*, 2014] included in this infrastructure is considered as a source of a newly created system. The Repository integrates metadata and full-text publications on geology of the Russian Far East, taken from territorially distributed different-type sources. At present the Repository contains more than 9000 metadata records and full-text articles in PDF format. The Digital Repository uses OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) and OAI-ORE (Open Archives Initiative Object Reuse and Exchange) to expose metadata.

Tables that are extracted from geological publications may have different structure and may be displayed in a variety of layouts. Matrix-like tables and nested tables are widely used in geological articles. Matrix-like table cells are laid out on a rectilinear two-dimensional grid. Nested tables enclose pictures, schemes etc.

In this article, we focus on the matrix-like tables. Such tables consist of three zones: title, note and elements of table (Figure 2). These zones provide important information about the names of geological ob-

jects, their geographical location, types of rocks and minerals, names of analytical laboratories, methods of analysis etc.

Publication Processing

Scheme of initial table processing in scientific publications is given in Figure 3.

Data Harvesting

The OAI-PMH allows to harvest all metadata descriptions of records from the digital repository, HTTP – full-text articles in PDF format, both protocols works in automatic mode.

Detection and Extraction

In order to implement the method of table boundary detection, automatic segmentation of bitmap images is used. PDF documents are converted into bitmap files with the help of a virtual printer. Blocks of columns constructed in accordance with the data retrieved after boundary detection are divided into cells based the layout of text blocks. The process of localization and

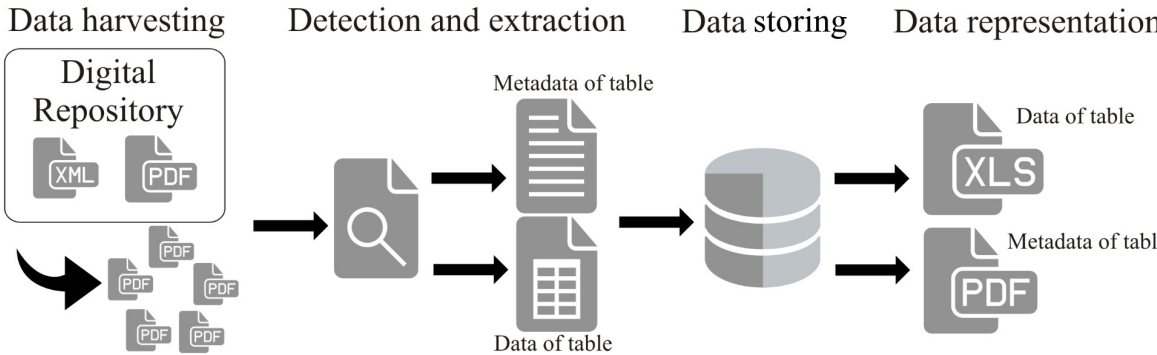


Figure 3. Conceptual scheme of initial table processing in scientific publications.

extraction of the table title and note is performed in the similar way.

Data Storing

Table data, descriptions, publication metadata are kept in the local data storage system run by the relational database management system (DBMS). Entity-relationship model (ER) describes the following main entities: publication, table, catalog, (subject), item (object). Part of the ER-diagram is shown in Figure 4. Database includes the following information: publication metadata record in XML format, table record, quantitative data table in XML format, list of objects for catalogs and a list of catalogs.

Data Representation

The System gives the possibility to search information by the use of catalogs and search engines. There are subject catalogs. The catalog of deposits includes more than 10,000 names from the list of Cadastre of Deposits of the Russian Federal Geologic Fund "ROSGEOLFOND". A user can use a simple search through standard fields of the document (authors, title, key words, journal etc.) or through zones of table record

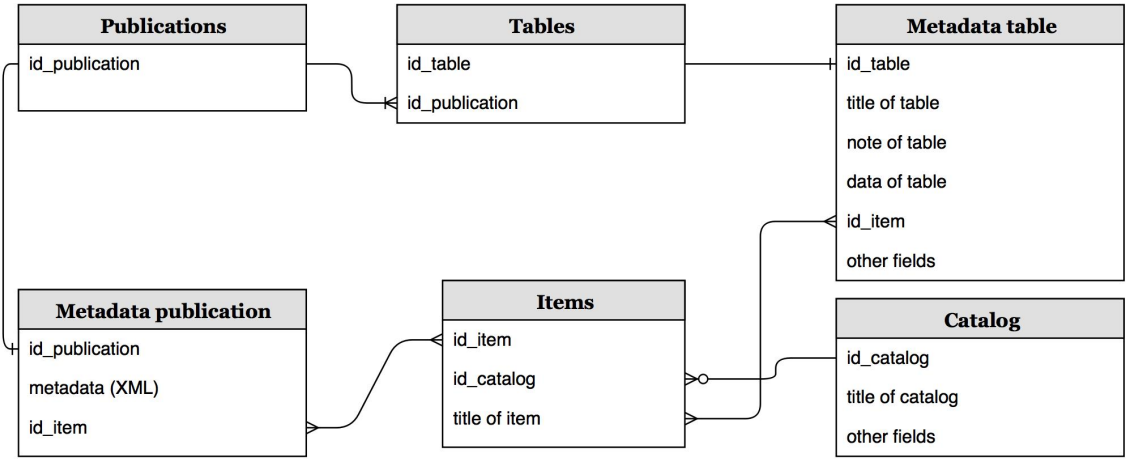


Figure 4. ER-diagram of the main entities.

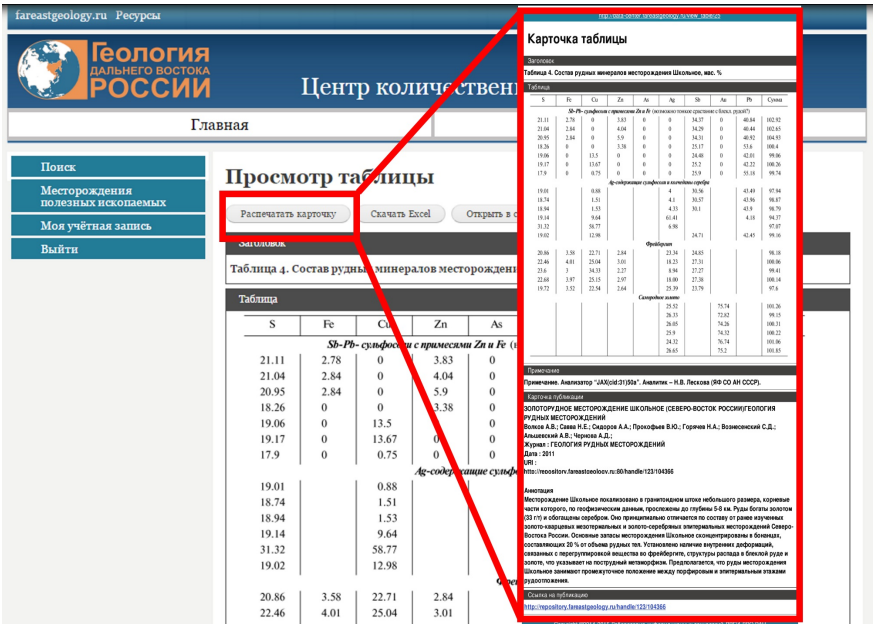


Figure 5. A sample of data provided to the user in PDF format.

(title, note etc.). By user's request the System forms a table record with output data in PDF format (Figure 5). The metadata record contains table title, bitmap image, table note, standard descriptions of publications and the reference to the original article in the Internet.

The user can save the retrieved quantitative data in Excel format (Figure 6). This format is widely used in geology as an effective tool of plotting charts and linear

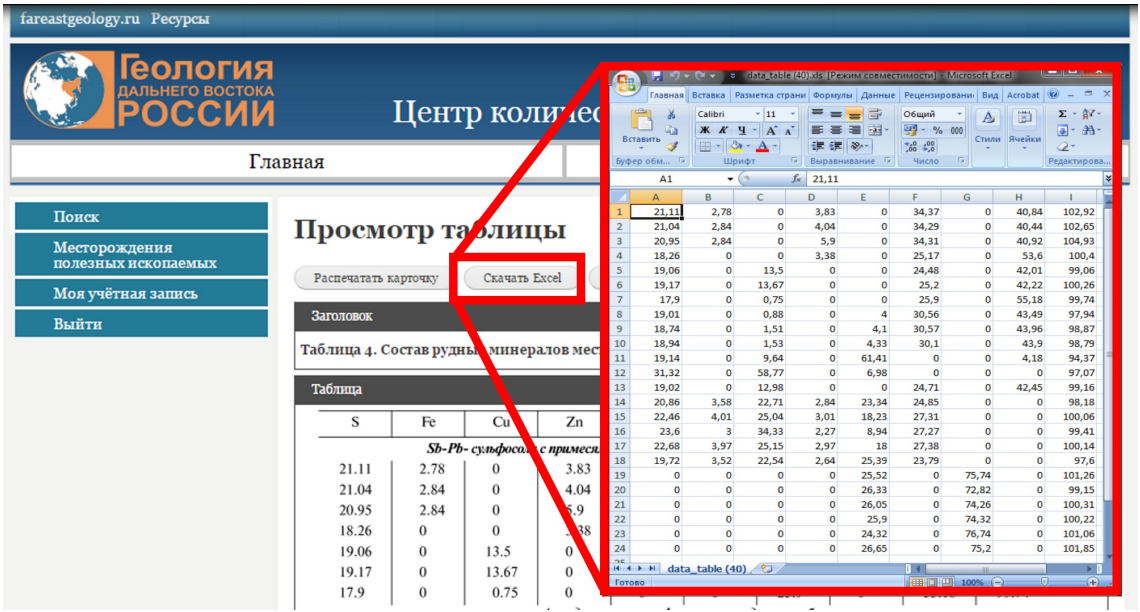


Figure 6. A sample of data provided to the user in Excel format.

dependence on the user's PC.

System Prototype

General functional scheme of the system is given in Figure 7. The scheme not only shows harvesting modules, access control system, user interface and admin interface, but also search, data processing, visualization and export services. Settings and correction of data inputting process in the System are carried out trough admin interface.

The Prototype of the system is available through the Internet. The Prototype is designed as a module of an open software platform CMS Drupal run on a general-purpose scripting language PHP. Drupal architecture makes it possible to use PHP and create information systems of any complexity. Default functions can be extended by adding supplementary extensions.

Acknowledgment. The work is supported by RFBR, No. 14-07-00068.

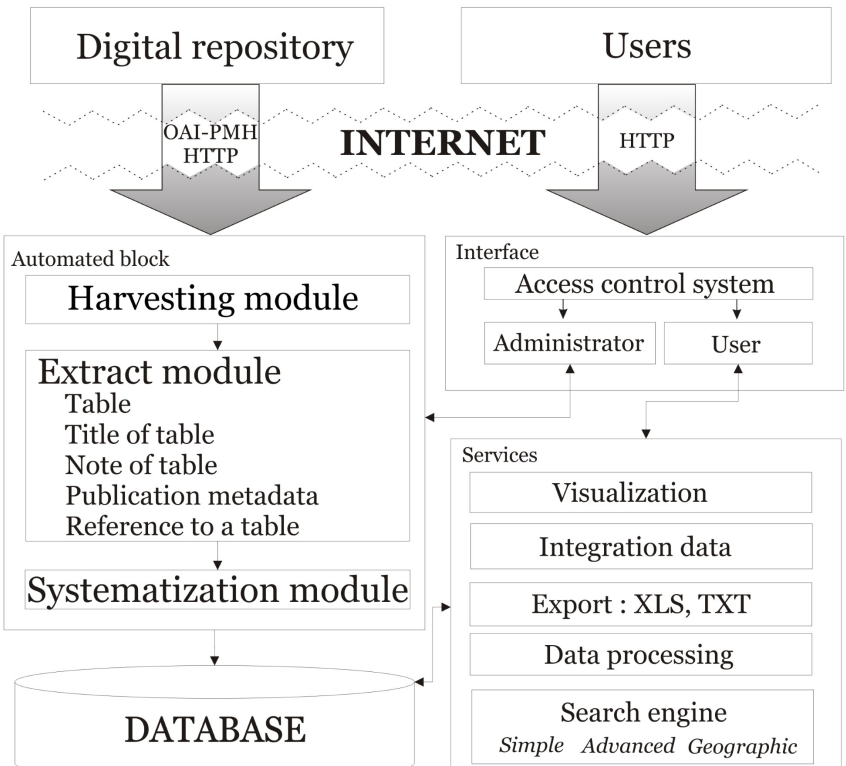


Figure 7. Functional scheme of the System Prototype.

References

Bart, E. (2012), Parsing tables by probabilistic modeling of perceptual cues, Document Analysis Systems (DAS), *10th IAPR International Workshop*, p. 409–414, Gold Coast, Queensland, Australia.

Day, R. A. (1979), *How to Write and Publish a Scientific Paper*, 160 pp., ISI Press, Philadelphia.

Downs, R. T., M. Hall-Wallace (2003), The American mineralogist crystal structure database, *American Mineralogist*, No. 88, 247–250.

Hassan, T., R. Baumgartner (2007), Table recognition and understanding from PDF files, *Proc. 9th International Conference on Document Analysis and Recognition (ICDAR 2007)*, IEEE Computer Society, p. 1143–1147, IEEE Computer Society, Washington.

Naumova, V. V., A. V. Belousov (2014), Digital repository Geology of the Russian Far East – an open access to the spatially distributed online scientific publications, *Russ. J. Earth. Sci., 14*, ES1004, doi:10.2205/2014ES000538

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

Sarbas, B. (2008), The GEOROC Database as part of a Growing Geoinformatics Network, *Geoinformatics 2008 – Data to*

Knowledge, Potsdam, p. 42–43, USGS, Reston.

Seo, W., H. Koo, N. Cho (2015), Junction-based table detection in camera-captured document images, *International Journal on Document Analysis and Recognition (IJDAR), 18*, 47–57.

Shigarov, A. O. (2009), Technology for Table Data Extraction From Digital Documents of Different Formats, *Thesis cand. tehn. sciences*, p. 143, ICT SB RAS, Irkutsk. (in Russian)

Tkaczyk, D., P. Szostek, M. Fedoryszak, P. J. Dendek, L. Bolikowski (2015), CERMINE: Automatic extraction of structured metadata from scientific literature, *Int. J. on Document Analysis and Recognition, 18*, 1–19.