Federal Information System on Grey Literature in Russia: a New Stage of Development in Digital and Network Environment

Aleksandr V. Starovoitov, Aleksandr M. Bastrykin, Anton I. Borzykh and Leonid P. Pavlov
Centre of Information Technologies and Systems of Executive State Authorities, Russia
19 Presnensky Val St., Moscow, 123557, Russia

Introduction

Since the late nineties when the Russian grey literature (GL) system in the sphere of scientific and technical information was first presented to the international GL community in Luxembourg we have had several opportunities to describe one or another aspect of the Federal Information System on GL in Russia [1,2,3,4,5]. This time we would like to dwell upon the system as a whole following its development from the past through present times to the prospective view all the more as this year a new ambitious project is started with the aim of renovating the system in accordance with the up-to-date requirements. The project has received a sufficient government funding for the coming three years.

The Russian Federation has inherited the federal-level information system on grey literature from the Soviet Union. The system covers the most informative kinds of grey literature - scientific research and development reports and post-graduate theses as the sources of scientific and technical information being centrally collected at the Centre of Information Technologies and Systems of Executive State Authorities (abbreviated in Russian as CITIS) in accordance with the Federal Law "On the obligatory copy of documents". The law obliges all the organizations – the collective authors of reports and persons – the individual authors of dissertations to give a free full-text copy of the documents to CITIS. In turn, the Centre is obliged not only to complete and permanently store the collection but also to disseminate the information on its content.

In the course of the past decades the system experienced several modifications in order to get adapted to the changing organizational and technological reality. In its present state the federal system combines the following three functionally separate systems run by CITIS: the traditional system for collecting, processing, storing and providing access to R&D reports and theses called the computerized information system on science and technology (abbreviated in Russian as ASINIT) that has recently been improved to store the full-text reports and dissertations in a digital form and provide full-text search and retrieval; the system for self-funded research projects registration and monitoring that was put into operation in mid-2000 to reflect a growing trend in funding R&D projects from research organizations' own financial resources; the federal register for the results of scientific and technical activities also created in mid-2000 with the idea of monitoring the life-cycle of patentogenic findings documented in scientific reports.

All the three systems are operative under the name "United Federal Database on Research and Development" (UFD R&D) and fulfil their functions however rapidly changing digital and network technologies create new environment to increase the systems' efficiency and improve its services. A new project in the process of development at CITIS is under the auspices of the newly-started State Programme of the Russian Federation "Information society (2011 – 2020)". The project is aimed at the creation of the Integral state information system on scientific research and development that is supposed to unite the three systems using unified forms of input documents so that users were to fill in the similar information only once and in interactive network conditions. The integral system will use the instruments of full-text digital documents analysis and web-technologies so that to improve data-mining and to avoid plagiarism.

The past

The computerized information system on science and technology (ASINIT) has been operating since 1975. It was then created as the grey literature part of the national library-information fund of the USSR and the part of the State System for Scientific and Technical Information (abbreviated in Russian as GSNTI). ASINIT consisted of two divisions: the full-text R&D reports and dissertations (the so-called primary documents) stored on microfiches and the bibliographic cards with abstracts (the so-called secondary documents) stored in the mainframe computer in a database format. There are two kinds of the secondary documents: the registration cards that are filled in when a new R&D project is started and the information cards that accompany full-text reports and dissertations.

Later on, in the early eighties ASINIT became the host core of the computer network called AIST in Russian for "computerized teleprocessing information network". AIST connected distant smart terminals, a prototype of personal computers situated all over the country, to the ASINIT host-computer with the grey literature databases situated in Moscow. The network operated in a dial-up mode through the public telephone lines. The distant users could conduct online searches in the centralized databases on reports and theses and order copies of documents from the System GL collection. The network throughput provided for more than 500 search, retrieval and copy-ordering transactions per day. That was the first information computer network of the pre-Internet epoch commercially working in the country.

No matter how obsolete the soft- and hardware of the System may seem now from the very beginning ASINIT met the main complex of requirements for completing the obligatory copy grey literature collection (R&D reports, candidate and doctoral dissertations – theses, descriptions of algorithms and computer programs), federal

registration of the documents, the database support, online search and retrieval, abstract journals publishing, permanent storing and archiving the documents [1].

This system's configuration existed for several decades with the technological changes from mainframe computers to PCs, database and network servers and the information migrating from magnetic tapes through diskettes and CDs to the modern electronic data stores.

The present

At present ASINIT is still the heart of the United Federal Database on Research and Development (UFD R&D) along with other two systems appeared several years ago. All the systems are functioning on the technological platform of the Centre of Information Technologies and Systems of Executive State Authorities (CITIS). In 2004 by the Decree of President of the Russian Federation ASINIT was included in the list of strategically important systems. Since 2010 the System has been listed in the Federal Register of the State Information Systems.

By the end of 2010 the system supported the following information resources:

- the retrospective bibliographic database with abstracts for R&D projects registration cards and R&D reports information cards containing nearly 2,5 million documents with the depth of retrospective 30 years (each card consists of more than 30 information fields) including
  - registration cards – nearly 1,2 million;
  - information cards – nearly 1,3 million;
- the retrospective bibliographic database with abstracts for dissertations containing more than 640 000 documents with the depth of retrospective 30 years (each card consists of 35 information fields) including
  - candidate dissertations information cards – nearly 560 000;

- doctoral dissertations information cards - more than 80 000;

- the abstract journals database – nearly 3,0 million documents;
- the database for information cards translated into English – more than 80 000 documents;
- the algorithms and computer programs database – more than 15 000 documents;
- full-text R&D reports (since 1984) – nearly 800 000;
- full-text dissertations (since 1984) – nearly 600 000 including
  - doctoral dissertations - nearly 80 000,
  - candidate dissertations – more than 500 000;
- the database for scientific organizations submitting R&D reports – more than 6 000 organizations.

The report and dissertation information cards databases are placed on a CITIS server with online network availability for the users. The databases serve as an electronic catalogue for the full-text collection providing a fast means of registration and search. The arriving full-text paper reports and dissertations are being scanned and PDF stored. At the same time the earlier documents are retroconverted (now backwards to the year of 2000) from the microfiches to PDF format. About 11 000 full-text R&D reports are entered into a full-text database. For the beginning of 2011 the total size of the electronic document store is 5 TByte. The total size of the information fund – more than 7 million documents.

The desk-top publishing system allows for issuing both electronic and paper abstract journals but now only the electronic versions are commercially disseminated by subscription. 51 titles of the journals by 25 subject series are published, totally 236 issues per year.

There are two government level documents which form a legal ground for the operation of the system: the Federal Law “On the obligatory copy of documents” of December 29, 1994 № 77-FZ (in the wording of March 26, 2008 № 28-FZ)

and the Government Decision of March 31, 2009 № 279 that delegated all the functions of running ASINIT to CITIS.

The system collects and controls scientific and technical reports and dissertations concerned basically all scientific subjects ranging from mathematics, physics, electronics and engineering through to social sciences and the humanities and supports monitoring and controlling the situation (both in financial and subject respect) in *the state funded* scientific research and development activities covering extensively all the territory of the Russian Federation [2,3]. The system's collection is an indispensable source for government agencies with an interest in the latest Russian contributions to science and technology.

At the same time it is evident that no matter how much money is given to science from the state budget it can never be the only and sufficient financial source for research and development and the diversification of funding is inevitable. So, there is a growing trend in scientific research that more and more R&D projects are being funded from *research organizations' own financial resources.* Those organizations are commercial ones functioning in the forms of federal state unitary enterprises and open joint-stock companies with the state share-holding. Their self-funded research projects were out of centralized monitoring and hence were not taken into account when updating the lists of priority development directions in science and critical technologies of the Russian Federation.

To eliminate the defects in research monitoring a special Government Decision was issued on November 4, 2006 No. 645 with the idea of creating a system for self-funded research projects registration [4]. The system was designed in the years of 2007 – 2008 and now is in operation as the second part of the United Federal Database on Research and Development (UFD R&D). Based on the output information from the system the Annual Summary Report for the Government is prepared. In accordance with the Decision the information on self-funded research

should be submitted in an approved unified form as an annex enclosed in the organization's annual financial report. The approved blank form is added to the Decision's text. The form's fields of data are important because their filling determines the information value of the document.

Now there is a four-year retrospective database (with the report documents of 2007- 2010 years – totally about 1,5 thousand documents), next year (2012) the documents of 2011 will be entered and so on. Thus, the system ensures the registration of report documents on self-funded research, their permanent storage in the database format and both quantitative and qualitative analysis of self-funded research in Russia prepared in the form of the Annual Summary Report. The self-funded research monitoring system is evidently grey because its input form and output Annual Summary Report are typically grey documents. Since 2010 the System has been listed in the Federal Register of the State Information Systems.

The grey literature sources contain a bulk of findings to be commercialized and/or claimed as intellectual property objects. The registration of reports and dissertations that is carried out in ASINIT now is rather document- than result-oriented. It would be useful to follow all the lifecycle of a scientific result beginning with the idea and basic research outcome through feasibility study findings to industrial implementation of the result in the form of innovative products and services [5].

In 2005 a Government Decision was issued (No. 284, now it functions in the wording of August 18, 2008 No. 622) on the development of the United Register for the Results of Scientific and Technical Activities (UR RSTA). In 2006 the Register was put into operation with its separate input forms designed to register the objects of intellectual property (patents, databases, computer programs, etc.) obtained in the course of state-funded research. Now the Register database contains the information on 50 ministries – the state R&D projects customers,

15 000 state contracts concluded by them to carry out the projects and 6 000 intellectual property objects. The Register is the third component of the United Federal Database on Research and Development (UFD R&D) operating in CITIS.

Though functionally satisfying the main requirements of the Law and scientific community the existing UFD R&D suffers from several shortcomings that are supposed to be eliminated in the course of the further system's development:

a) all the databases (DB) on R&D — state contracts DB, reports and dissertations DB, the Register DB – are formed independently one from another, so the user has to fill in several similar forms wherein the information is redundant and duplicated;

b) the lack of effective customers and executors control mechanisms, so to say, a feedback from the System to the customers in order to provide the completeness of R&D reports registration and presence in the System;

c) the total computer power of the System is insufficient to implement the modern web-technologies of forming the information resource and providing a comfortable access to it;

d) the limited analytical means of textual information processing;

e) there is no online interaction with other state information systems such like the Computerized information system for scientific research of the Russian Academy of Sciences.

The future

The newly-started State Programme of the Russian Federation "Information society (2011 – 2020)" has opened a real chance for the state financial support of the System's development in the context of rapidly changing digital and network technologies. Under the auspices of the Programme a competition was announced for a state contract to realize the System's development project. CITIS won the

competition and the contract was concluded for three years to fund the project named “The development of the United R&D projects registration system (UPRS R&D) for the projects carried out in the civil sphere with the state budget funding”. In general, the project is aimed at the creation of the integral state information system on scientific research and development that is supposed to unite the three systems functioning on the platform of CITIS.

There are some main problems to be solved within the project:

- The development and implementation of effective mechanisms for securing the information completeness in the System that is all the R&D reports must be registered and present in the System. This is very important because the experience of the latest decades exposed a low executive discipline of scientific and scholar institutions that perform R&D projects.
- The elimination of redundancy and duplication in the inputted and stored information due to the existing database conducting independence. Different databases have different forms of records with many coinciding information fields.
- The registered data must include not only the subject of research information but also the data on the state contracts, size and structure of the state funding, patentogenic results of the research project.
- The development of analytical instruments to expose the innovating projects and estimate the results of conducted research.
- The development of the legal basis for the UPRS R&D. A new Government Decision regulating the procedures of the System’s operation must be prepared and approved.

From a technological point of view the system’s modernization must develop in the direction of network computing and digital documents processing. The essential points of novel approaches are the following.

1. The unified forms of the secondary documents – information cards – are developed so that on the one hand to eliminate the duplication of the

same fields in different cards and on the other hand to include more detailed financial data on the size and sources of funding and data on the life cycle of the intellectual property objects (patents, computer programs, databases, etc.).

2. The online mode of filling the new forms of information cards is provided for the authors of R&D reports and dissertations who are able to address the CITIS site on the Internet (www.rntd.citis.ru), click "the online form filling-in" and have the form on the screen of their computer. There are many conveniences supporting the online filling-in such as the formal verification of numerical fields, the enclosed lists of priority directions and critical technologies and the list of correct names of the organizations that were previously registered in the system. The user just has to click the name instead of keying it in.
3. The formation of digital full-text databases for all the arriving documents (reports and dissertations) with the effective means of full-text search and analysis. The digital documents are entered into the single electronic repository that allows four modes of documents entering: scanning and recognizing the paper documents; inputting the documents arriving on CDs; online arrivals entering; retroconversion of the documents stored on microfiches. Now, in accordance with the existing legal acts, the full-text documents arrive on paper and must be scanned and digitized before being PDF stored. The evident tendency is to pass on to digital input documents.
4. In order to introduce exclusively digital input forms of both the full-text and metadata documents it is necessary to implement an electronic signature technology and to make alterations in the legal acts (laws etc.) currently in force.

There are two kinds of subsystems envisaged in the technical assignment for the new system: those existing and being modernized and those newly designed and implemented.

The modernized ones are:

- the subsystem for reports and dissertations collecting, processing and registration;
- the digital documents repository and archiving subsystem;
- the search and retrieval subsystem;
- the abstract journals publishing subsystem.

Among the newly designed ones are:

- the system's common Internet portal subsystem;
- the R&D projects in progress monitoring and content analysis subsystem;
- the subsystem for interaction with international scientific and technical information systems;
- the subsystem for integration with other Russian state information systems on science and technology.

In the framework of the integral system a new complex of analytical and search instruments is to be designed using artificial intelligence technologies for linguistic text processing and semantic analysis, context and fuzzy search algorithms, subject area structuring, new knowledge and data-mining, antiplagiarism and experts activity support. This will allow to create analytical information not only about a separate scientific work but also about scientific trends, scientific groups and schools, the information for updating and systematizing scientific classification schemes. A linguistic support of these possibilities suggests that computer glossaries and dictionaries, thesauri, ontologies and classifiers should be developed and maintained.

Concluding remarks

During the next three-year stage of development it is supposed to implement the advantages of digital and network technologies and significantly improve the system's characteristics. The system is designed to provide a complete R&D documents collection, a fast access to full-text documents and relevant information. It will allow to monitor the situation in the sphere of R&D works and projects all over Russia, to support the federal level administrative decisions in the sphere of science and technology, to prognosticate its development, to improve the distribution of financial means for scientific R&D, to reduce the unjustified duplication and overlapping of R&D projects and dissertations.

References

1. Pavlov, L.P. The State and Development of the Russian Grey Literature Collection and Dissemination Centre. – "Interlending and Document Supply". - MCB Univ. Press. 1998, vol. 26, No. 4. Pp. 168-170.

2. Pavlov, L.P. Literatura gris rusa en un mundo digitalizado e informatizado. - "Ciencias de la Informacion." – La Habana: IDICT, 2002, agosto, v.33, N 2, pp. 39-42.

3. Pavlov, L.P. Legal Foundations of the Scientific and Technical Grey Literature Development in Russia. - "The Grey Journal". Internat.Journal on Grey Literature, Amsterdam. Spring 2007,vol. 3, N 1, pp.37-43.

4.. Starovoitov A.V., Bogdanov Yu. M., Bastrykin A. M., Pavlov L.P. The Grey System for Monitoring Self-Funded Research. - GL11 Conf. Proc.– GL-conf. series, N 11. 11th Internat. Conf on Grey Literature, 14-15 Dec.2009, Washington D.C.- Amsterdam:TextRelease, Dec. 2009.- 142 p. P.13.

5. Pavlov, L.P. The Commercialization of Research Findings Documented in Grey Literature. - Proc. 5th Internat. Conf. on Grey Literature: Grey Matters in the World of Networked Information. – 4-5 December 2003 Amsterdam/ GreyNet.- Amsterdam: TextRelease, January 2004.VI. – Pp.64 -68.