SOFTWARE DEVELOPMENT FOR DISTRIBUTED SYSTEM OF RUSSIAN DATABASES FOR ELECTRONICS MATERIALS

Valery Kornyshko, Victor Dudarev

Abstract: Current state of Russian databases for substances and materials properties was considered. A brief review of integration methods of given information systems was prepared and a distributed databases integration approach based on metabase was proposed. Implementation details were mentioned on the posed database on electronics materials integration approach. An operating pilot version of given integrated information system implemented at IMET RAS was considered.

Keywords: distributed database integration, metabase, Web services, database on electronics materials.

ACM Classification Keywords: C.2.4 Distributed applications, Distributed databases; D.4.4 Network communication.

Introduction

Development and utilization of databases for substances and materials properties is a basis in providing information service for specialists in chemistry and materials science. Every research organization aimed at its own data center creation. Such data centers contain information closely related to research areas of a particular organization. Historically several data centers were formed for data storage and processing in every organization (scientific research institute or university). This can be explained not only by administrative reasons, but rather by significant differences in the problem domain. Existing situation creates great problems for accessing such data, because this information is dispersed over numerous data sources.

At present time, period of such an information fragmentation is coming to the end due to rapid IT-industry development. Present-day progress in science and technique stimulates concentration of diverse information on physicochemical substances properties. Modern polyfunctional materials development requires from us high standard of knowledge in different properties of substances. Efficient online information service (for materials science engineers and chemists providing full data from reliable sources) decreases baseless papers' duplication and ultimately it reduces cost and time required for modern materials development. Inaccessibility and frequently dispersion of information over different heterogeneous data sources makes great difficulties in decision-making process considering application of one or another material.

During development integrated information system presented in this paper the key task was to create an intelligent, simple in architecture and effective software infrastructure. This software infrastructure should integrate data on properties of substances and materials rationally and reasonably. Integration means are required that should be capable to provide not only unified access to operating data centers, but these integration means should allow us to create comprehensive data access infrastructure based on unified standards and also on uniform network interconnection principles.

Database Integration Approaches Overview

Principally there are two approaches to database integration.

The first one implies full merging of existing resources. That is the case when database complex is a single information system (megabase) for end users, operators and administrators. Database exploitation costs reduction and information duplication decrease can be mentioned among advantages of this very variant.

Every data center is a point of information concentration and online data analytical processing. In addition, technology of information accumulating and data processing is settled down in each organization. Moreover, great investments that were made in hardware and software do not allow solving the data dissociation problem by mechanical transportation to some centralized database of all data. Moreover Russian databases for electronics

materials have been developed in various organizations and thus they took advantage of different database management systems (DBMS). Taking into consideration differences in data quality, data expertise, data store types and many other troubles emerging when changing existing systems operating principles it should be stated that full and smooth integration is practically impossible for above mentioned resources.

The second integration approach main essence is that we are not going to integrate databases themselves, but we want to integrate their proprietary user interfaces only. From the one hand this approach allows us not to change every integrated database structure dramatically (and thus established database utilization and administration technology – data update and insert). From the other hand, this approach allows the end user to get access to the whole information picture on chemical substances stored in different databases. So called "virtual" database integration (or in other words heterogeneous information system creation) implies independence in evolution of separate subsystems and at the same time end user gets access to the whole information array on a particular chemical substance or material stored in databases of a virtually united system (fig. 1). And that fact solves the main integration goal truly.

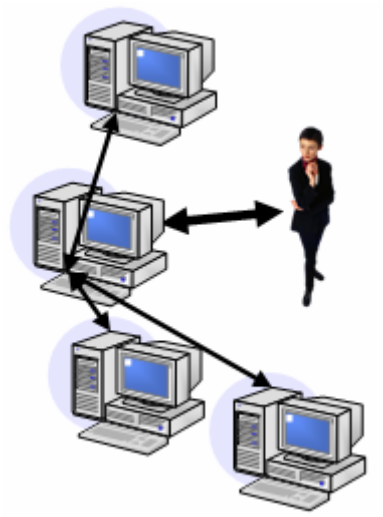

Fig. 1. Database integration at Web interface level

Taking into consideration current development conditions of Russian databases on physicochemical substances' properties the second integration approach – integration at interface level only – is more appropriate and quite perspective. It's worth mentioning that Web-interfaces have been developed for IMET RAS databases on physicochemical substance properties ("Crystal" database on acoustic-optical, electro-optical and non-linear optical substances and "Diagram" database on semiconductor systems phase diagrams). These Web-interfaces allow users to get remote access via Internet to data stored in these databases using any Web browser.

Searching for Relevant Information in Integrated System

When integrating databases at Web-interfaces level it's required to provide facilities for browsing information contained in other databases. This information should be relevant to the data on some chemical system currently being browsed by user. Let's consider this in the following example. User who browses information on Ga-As system from "Diagram" database should have an opportunity to get information for example on piezoelectric effect or non-linearoptical properties of GaAs substance contained in "Crystal" database.

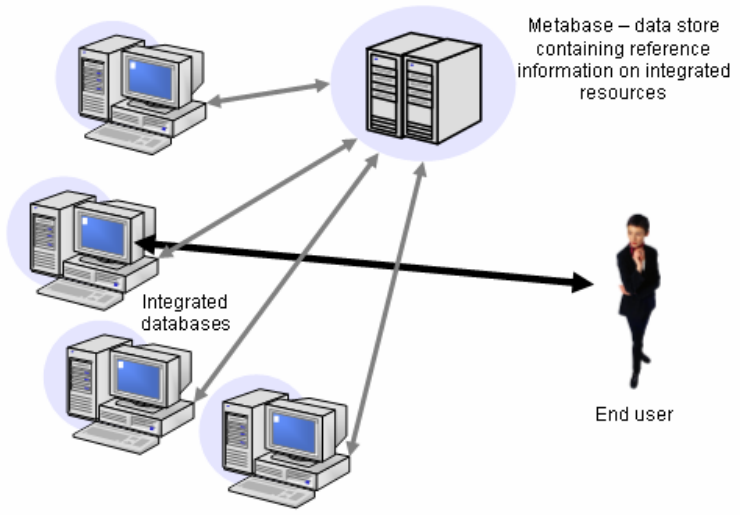

Fig. 2. Metabase concept

So it's obvious that when designing distributed information system, it's required to provide search for relevant information contained in other databases of distributed system. Thus, we hardly need to have some active data center that should know what information is contained in every integrated database. Obviously some data store should exist that somehow describes information contained in integrated database resources. In this manner, we come to the metabase concept – a special database that contains some reference information on integrated databases contents (fig. 2).

In our case, it is information on chemical systems and their properties. The amount of this metainformation should be enough to perform search for relevant information on systems and corresponding properties.

Let's try to formalize the problem in terms of set-theoretic approach. Hence, metabase should contain information on integrated databases (D set), information on chemical substances and systems (S set) and information on their properties (P set). To describe correlation between elements of D, S and P sets let's define ternary relation called W on set $U = D \times S \times P$.

Here U is a Cartesian product of D, S and P. Membership of a (d, s, p) triplet to the W relation, where $d \in D, s \in S, p \in P$, can be interpreted in the following way: "Information on property p of chemical system s is contained in integrated database d ".Having defined three basic sets it can be seen that search for information relevant to s system can be localized to determination of R relationship, that is a subset of Cartesian product $S \times S$ (or in other words, $R \subset S^2$). Thus, it can be stated about every pair $(s_1, s_2) \in R$ that chemical system s_2 is relevant to the system s_1. So all we need to solve the task of searching for relevant information in integrated databases is to determine somehow the R relation. It is significant to note that R relation can be created or complemented by means of either of two variants. The first variant is via using predefined rules by a computer. The second one is that experts in chemistry and materials science can be engaged to solve this task.

The second variant is quite clear – experts can form relationship R manually following some multicriterion rules affected by their expert assessments. So let's consider possible variants of automatic R relation generation. One of such variants can be like this one based on the following rules:

1. For any chemical systems $s_1 \in S, s_2 \in S$ composed from chemical elements e_{ij} $s_1 = \{e_{11}, e_{12}, .., e_{1n}\}, s_2 = \{e_{21}, e_{21}, .., e_{2m}\}$ it is true, that if $s_1 \subseteq s_2$ (i.e. all chemical elements of system s_1 are contained in system s_2), then $(s_1, s_2) \in R$.
2. R relation is symmetric. In other words for any $s_1 \in S, s_2 \in S$, it is true, that if $(s_1, s_2) \in R$, then $(s_2, s_1) \in R$ as well.

These two rules allow us to determine a set of chemical systems relevant to the given one. It should be noticed that this automatic R relation generation variant is just one of the simplest and most obvious variants of such rules, and in fact more complex mechanisms can be used to get R relation. For example, browsing information on a particular property of a compound in one of integrated databases (in fact, it is information defined by (d_1, s_1, p_1) triplet), we consider (d_2, s_2, p_2) triplet to be relevant information. (d_2, s_2, p_2) triplet characterizes information on some other property of a system from another integrated database. In this case, we have got more complex relevance relation like this $R \subset (d_1, s_1, p_1) \times (d_2, s_2, p_2)$, where $d_1, d_2 \in D; s_1, s_2 \in S; p_1, p_2 \in P$. In fact we can even define a set of several R relations ($R_1, R_2, ..., R_n$) by applying different rules. Thus we'll be able to perform search for relevant information based on wide variety of R interpretations.

Loading Information into Metabase

As it was stated above, Russian databases on materials science have been developed in different organizations on various platforms and that fact makes integration significantly more complicated. Metabase should store reference data on integrated resources contents. It's obvious that in this situation it's required to use open network interconnection principles and standards supported on multiple platforms. If we consider present technology stack then it's quite clear that currently Web services are connection links between different platforms and heterogeneous environments. Web services are based on common standards such as SOAP (Simple Object Access Protocol) and XML (eXtensible Markup Language). Nowadays these technologies are capable to provide reliable infrastructure for cross platform message exchange.

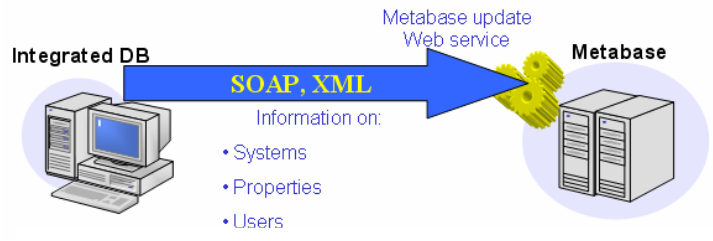

Fig. 3. Metabase update Web service

In that way reference information loading into metabase was implemented by means of metabase update Web service, so-called MUService (fig. 3). Let's consider metadata updating mechanisms in detail. System that is to be integrated with others should generate XML document which contains information on updates in that very system. The layout format of this XML document is generally standardized for all integrated subsystems and it is strictly fixed by means of specially developed XML schema [1]. Thus, all subsystems being integrated should generate valid XML document that meets XML schema requirements to notify metabase of information changes that occurred in their state. After being generated, XML document is being sent to the metabase update Web service for processing and metabase update. Interaction with this MUService is implemented by means of SOAP protocol according to the Web service WSDL (Web Services Description Language) description [2]. In that way client databases report about updates to the metabase and so actual information on integrated resources contents appears in the metabase.

It's important to note that security issues were among primary concerns while designing and implementing the resulting system. In that way symmetric encryption mechanism was implemented with the aid of DES (Digital Encryption Standard). It guarantees secure metadata exchange with metabase update Wes service. Additionally an option for data archiving was included into the system. A kind of zip-achieving was implemented that allows us to package data transmitted to the metabase server. This feature allows dramatically decreasing data volumes being transmitted via public networks (taking into consideration a high level of compression for XML documents). Thus this feature lowers requirements to network bandwidth and it is very important and actual for Russia since high-speed Internet access is not available everywhere in the country. Compressing techniques enable us to decrease information volumes so that it becomes possible to use old-fashioned data modems on telephone wires to transmit data.

As it can be seen, metabase update Web service supports some rather complex additional data transformations (encryption and archiving) requiring some extra coding. So to simplify the interaction process with the Web service a special Web client has been created. It was implemented as a COM object and thus it can be easily accessible from any environment which supports Microsoft COM. Created Web client addresses issues connected with encryption and compression of information that is to be sent to the Web service. It controls all network interconnection aspects also. All this functionality just simplifies routine database attachment to the integrated system.

It should be mentioned that at present time only integrated database systems (as client systems to the metabase) could initiate data update process with metabase update Web service. This technique of course is not the only variant of interaction scheme. Thus it is planned to redesign metabase update mechanism to enable metabase to inquire integrated resources on demand and thus to query information updates occurred.

It should be mentioned that after every metabase update session incremental population crawl is started on the metabase to update or to reindex relevant chemical systems list regarding information changes. This allows metabase to maintain actual information on relevancy relations of chemical systems contained in integrated resources. Currently relevant system reindexing is performed by means of approach of two rules proposed in this paper earlier. If it is necessary, these rules can be easily modified. And the main advantage is that in this very case there is no need to redesign the whole system concept. All we have to do is just to write a new piece of software to provide a new method of searching for relevant systems and replace the old module with the new one.

Metabase Integration – How It Works

Let's consider the operating process of the integrated system from the end user's point of view. In a general sense, the integration of information resources of materials science is in consolidation of available Web applications serving users of different materials science databases. This consolidation is provided by means of

specialized software but user should not be aware of it if possible. The software should be transparent in this sense.

When designing the integrated system special attention was paid to security system development. It should be mentioned that every developed information system has its own proprietary security facilities protecting the system and giving permissions to access it. Security system of a particular information system is responsible for granting permissions to the registered users of a given system only. It's obvious that in the context of integrated security system authorized users should have permissions required to get access to the information in integrated resources within their privileges strictly.

For example, let's consider the possible user work session scenario. A user has been granted access to database on semiconductor system phase diagrams "Diagram" and currently he or she is browsing information on In-Sb system. Obviously the user should have an opportunity to get information on elastic constants of In-Sb system from "Crystal" database. But that user should not be granted privileges to observe information on chemical systems other that In-Sb since he or she is not a registered user of "Crystal" database. And vice versa, if the user is a registered user of "Crystal" database too then he or she will be granted full access to "Crystal" as integrated resource. From our point of view, the described approach is an appropriate one and so it is used to design the distributed security system of integrated databases. It should be mentioned that user credentials of every integrated resource are also transmitted to the metabase via MUService as well. It is done to organize distributed security system operation in cooperation with corresponding security systems of integrated resources. It should be emphasized that open user passwords are not transmitted to the metabase, instead of open passwords, password MD5 hashes are transmitted in fact. This substitution (MD5 hash instead of open password) allows integrated information system to authenticate active user and at the same time this technique excludes possibility of using open user password to login to the integrated system database. In other words, there is no place for vulnerabilities here. So even if this data are stolen integrated resources can not be compromised.

Let's assume that in one of integrated system a user browses information on some particular chemical system. In other words, the user is in Web application of a particular information system (fig. 4). If it is necessary to get relevant information this Web application is capable to send a request to specially developed Web service [3] that serves users of integrated system. The request aim is to get information contained in integrated resources that is relevant to the currently browsed data. After the request the Web service sends a response to the Web application in a form of XML document. It describes what relevant information on chemical systems and properties is contained in integrated resources. As it was mentioned, earlier data in XML format are properly understood on all major platforms. That information can be output to user for example by means of a XSL-transformation in form of HTML document (XML + XSL = HTML) containing hyperlinks to special gateway. The user can follow from one Web application to another to browse relevant information via this gateway only.

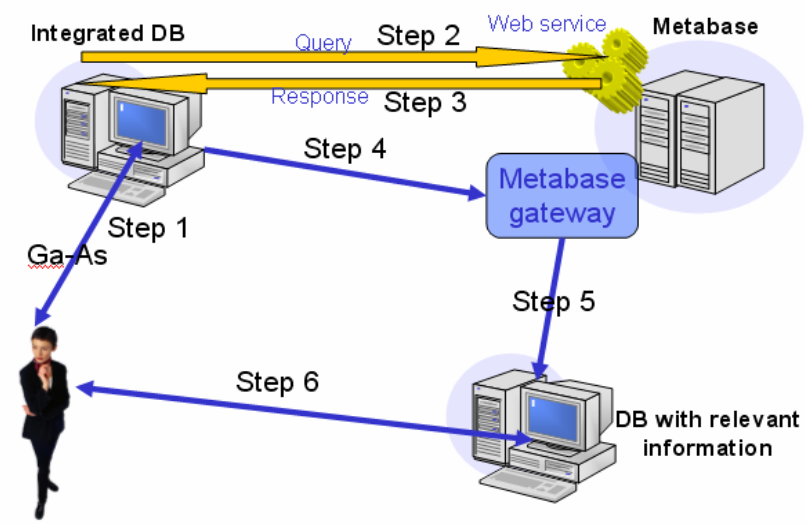

Fig. 4. Metabase integration – how it works.

Imagine that the user clicks on a hyperlink to start browsing information from some other integrated system. First of all, when the user has clicked the hyperlink, he is forwarded to the special gateway. Actually it is a specialized Web application that runs on the metabase Web server. The gateway main purpose is to perform security-dispatching function in distributed system. According to the task stated it is responsible for user authentication and it also checks whether the user has required privileges to address the information requested.

Let's imagine that authentication is successful and the user is eligible to address the data so the metabase security gateway performs redirection to a specialized entry point of desired Web application adding some additional information to create proper security context and a kind of digital signature. It should be stated that the entry point is a specialized page in target Web application that is to perform service functions for integrated system users. At this very page target Web application checks digital signature of the metabase security gateway and if everything is ok the page creates special security context for user with given access rights within target Web application. Finally, the user is automatically redirected to the page with the information required. In spite of redirection process apparent complexity, user transition from one Web application to another is absolutely transparent. Thus, end user can even not note that some complex processing has been done to perform redirection. So, it is an illusion created that having clicked on a hyperlink the user is simply transferred from one information system to another.

Conclusion

It's high time to draw a conclusion. The proposed database integration approach based on metabase was successfully applied at A.A. Baikov Institute of Metallurgy and Materials Science of the Russian Academy of Sciences (IMET RAS). "Crystal" and "Diagram" databases were the very first systems connected to the metabase integrated solution. This fact allows users of either information system to browse information from these databases. Now several words should be said about the project perspectives. First perspectives are connected with the resulting system extension due to addition of already developed Russian databases on materials science: IVTAN and MITHT databases. This integration will allow creating distributed database complex on electronics materials that has no analogs worldwide at present time. Besides numerical growth of integrated system there are plans for functional capabilities extension i.e. qualitative leap is planned. For example, it is projected to provide capability to perform complex distributed database queries that allow searching for substances that satisfy some defined complex criteria while information on criteria values is distributed over several databases. Consequently, to successfully perform such query it's required that metabase information system has an opportunity to query distributed databases impersonating acting user who initiates the initial complex query. After that the metabase should gather information from several sources, process it and output to the end user. Integration at that level undoubtedly will expand distributed information resources capabilities significantly.

Bibliography

[1] XML-schema that standardized XML document format for metabase update Web service is available at http://meta.imet-db.ru/MUService.xsd

[2] WSDL-contract that defines methods that can be utilized to interact with metabase update Web service is available at http://meta.imet-db.ru/MUService/MUService.asmx?wsdl

[3] WSDL-contract that defines methods that can be utilized to interact with Web service that serves integrated resources is available at http://meta.imet-db.ru/Service/Service.asmx?wsdl

Authors' Information

Valery Kornyshko – MITHT, Head of IT department; 119571, pr. Vernadskogo, 86, Moscow, Russia; e-mail: inftech@mitht.ru

Victor Dudarev – MITHT, junior member of teaching staff of IT department; 119571, pr. Vernadskogo, 86, Moscow, Russia; e-mail: vic@osg.ru