Design and Development of an Integrated Research Information System

Jānis Eiduks[1], Ainārs Auziņš[1], Gunārs Lācis[2], Inga Moročko-Bičevska[2]

[1]Riga Technical University, Riga, Latvia
[2]Latvia State Institute of Fruit - Growing, Dobele, Latvia

Abstract

Information systems' (IS) requirements are growing rapidly. The transaction management IS are realizing the simple data input/output process. The information management IS and the executive management IS are performing functions of the simple data analysis. The On-line analytical processing system (OLAP) uses multidimensional database (DB) form for extensive analysis and forecast. Recently another new IS class - the research IS has started to develop. Technical requirements for these systems are very high. The research ISs functionality is very broad, including different data loading and transformation, simple and complex calculations, use of rules and multiple criteria in decision-making. The following complex multifunctional IS building is really complicated, and in the existing realizations unique solutions often are used, but these solutions are usually expensive, time-consuming and otherwise limited. The user interfaces with the systems are complex, too - because of using many commands for data input/output, as well as language. In designing and developing research IS for fruit-tree biological resources investigation, there had been analysed researcher information processing behaviour, as well as investigated the latest IS and database system (DBS) technologies and concepts. These activities helped to work out dynamic, simple and effective IS common architecture, which provides high quality solutions for complex and complicated research tasks in a simple way.

Keywords: *Research information system, Information processing behaviour, Information processor, In database processing technology, In database analytic technology, Semantic in database technology.*

1. Introduction (research information systems)

It is necessary to realise many different tasks for carrying out the research work. Big amount of data, different data formats and many processing methods are used. All research IS elements for data storage, processing and communication organize a net structure. [1] Such structure realization is hard problem. A different type of research IS are designed and developed:

1. Laboratory IS (LIS) ensure receipt of data and information storage, processing and analysis for operation processes [6]. These systems often should cooperate with the different types of devices and other IS. LIS are integral systems or forms a part of some integral systems. LIS systems also are often called as "laboratories management systems (LMS)" or "laboratories information management systems (LIMS)". [2] The main functions of LIMS are: work flow tracking and locking, the data flow tracking and locking, the data exchange between equipment, computers and applications. For carrying out these functions, laboratory informatics and laboratory automation methods are used. [2] One of the most popular and experienced companies in developing LIMS systems is LABVANTAGE. It has developed a wide range of laboratory research systems, as well as created a common IS architecture, which is used in many other LIMS development companies. The LABVANTAGE LIMS systems are widely used control in meta- data level. [3]

2. Current research IS (CRIS) store data about the organisation research activities, publications, reports and results. In these systems IS standards of the European Union for CERIF (Common European Research Information Format) widely are used. [4]

3. **Scientific data management systems** (SDMS) focus on the exploratory data analysis and efficiency enlargement for automatic formation of documentation. [5]
4. **E - research IS** and **process development execution systems** (PDES). The accent in these systems is directly to the organisation of research cooperation. [6]

Unfortunately, existing research systems do not ensure fully the necessary broad functionality and simplicity in user's activity during task or related task chain. The use of rules and knowledge is underdeveloped, too. Existing designs are unique projects with a specific architecture and there is hard to make any improvements and extensions.

2. Use of cognitive psychology research models to define common information systems architecture requirements

In order to prevent the deficiencies of existing research IS, more attention should be drawn to human activity and behaviour during the research process. Human activity and behaviour in research is a cognitive process. Therefore, cognitive psychology concepts, theories and models could be utilized in IS architecture design. Most often used model of cognitive behaviour is "information processor" (see Figure 1. (a)) [8], but for practical needs it is to simple. It needs to extend with constructive ideas from other models.

Sternberg's "theory of intelligence" models cognitive process of three components set: creative, analytical and practical abilities components. [7] Creative component generates ideas, while analytical and practical abilities help to evaluate realization possibilities in order to verify the correctness of any idea.

In „working memory" model Sternberg expresses the phenomena of hierarchy of memory: sensor or perception memory, cognitive short - term memory and long-term memory. [7]

Piaget's "cognitive development theory" is based on human development and growth. According to human development there are four thinking stages identified: sensorimotor or reflex memory, simulation or preoperational memory, logic's use and compliance of various factors of memory (concrete operational memory), as well as abstract concept organization and arguing memory (formal operational memory) [7].

Including „intelligence theory", „working memory" and „cognitive development theory" ideas in starting „information processor" model, we expand it (see Figure 1. (b)). New model provides much more extensive and constructive understanding of the researcher's behaviour and activities.

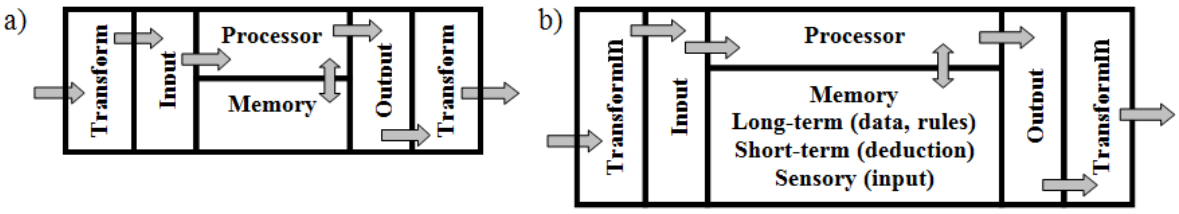

Figure 1. „Information processor" model (a) and extended „information processor" model (b).

3. New information systems' building technologies and conceptions

To create an effective research IS of good quality, there should be noted that in the last decade several new very important concepts and technologies have appeared, such as „big data" concept, „in database processing"

technology, „in database analytics" technology, „semantic in database" technology, „cloud computing" and „in memory database" technologies (these two technologies concern more to computers technical organization problems and are not discussed). [9]
Firstly, there exists a lot of „valueless" and unnecessary information (data duplication, bad storage structure). To solve this problem, it is important to create information filters of a high quality.

Secondly, serious work with data systematization should be recommended. Instead of big data amount it is important to use models and knowledge presentations, thus significantly reducing the data problems and improving the speed of resolution - in particular, in research IS.

3.1. Use of "in database processing" technology

The DB concept has experienced different evolutionary and revolutionary changes. Today the productivity of DB servers is very high. It does not need all power for data storage and extraction's organization. Part of abilities may be used for data processing tasks with intensive DB data use. It reduces network load and the run time really substantially. In the 1990s the beginners of realization of the concept were such companies as SAS, IBM, Oracle, and Illustra/Informix. Two important factors appeared which contributed the data pre-treatment task inclusion in the database server.

1. In 2010 the query language SQL (Structured Query Language) Standardization and development committee SQL-09 simplified query syntax to bring nearer SQL language to natural everyday language. This allows creating easier perceptible and applicable communication forms with DBS for users (languages New SQL „Jdb", New SQL „S2") with DBS. The user can retrieve necessary data in "ad hoc" situations by defining in simple and intuitive way his needs in new SQL language.

2. Universal DBS now have various ways of data storage. There are used relational, object – relational, XML and graphical DB storage formats.

In addition, it is possible to include new types of data and provide the necessary processing functionality. DBS also realizes mutual transformations of all data types. It is possible to use virtual DB for retrieval data in necessary format, not only in the format in which data had been stored.

These factors were the basis for the new DBS technology – "in database processing" starting. It could be mentioned that additionally essential factor that influence "in database processing" technology implementation is multiple programming language use in DBS for stored procedures programming.

Interesting and effective innovation of technology "in database processing" is attached systems. In DBS are included direct link with the spreadsheets. This allows you to keep your data in the database, output them in spreadsheet, carry out pre-calculations in spreadsheet; input data back to DB and perform big calculations in DB.

3.2. Use of „in database analytics" technology

DBS must include different data analysis methods for the research purpose and these methods are not only a simple calculation. Therefore, DBS include multidimensional DB, data mining technology and on-line analytical processing methods. This means that the user can create virtual multi-dimensional DB from the existing data and work with it. The data processing transaction is created with the data mining methods in the same way. The user retrieves the necessary data from the DB, defines parameters for the analysis model and initializes the execution of the method. Such possibilities have obtained the term "in-database analytics". [10]

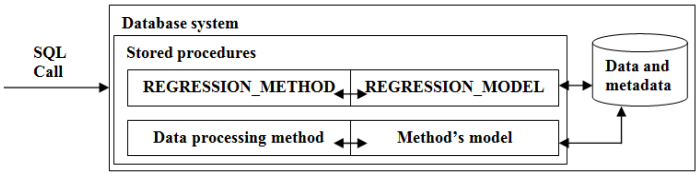

Figure 2. Processing methods and their additional models.

Such DBS analysis methods as stored procedures can be performed with the SELECT queries, but more complex calculation methods can be realised using so called "model construction". The parameters of the method and instruction storage data structure (model) are created (see Figure. 2). The execution sequence is following:
1) insert into DATA_FOR_REGRESSION select … from ... where ... (Organization of needed data.)
2) insert into REGRESSION_MODEL values('DATA_FOR_REGRESSION', (Input parameter values and 'A[1]+A[2]*X[1]+A[3]*X[1]*X[1]-A[4]*X[2]*x[2]'); indications.)
3) select REGRESSION_METHOD('REGRESSIN_MODEL') from DUAL; (Execution of method.)
4) in the result values of regression function coefficients are obtained. → A[1] A[2] A[3] A[4] .

3.3. Use of "semantic in database" technology

DB systems' semantic technology is designed for searching and processing information in the Semantic web, but it can also be used locally only for DB requirements. Semantic technology in DBS is based on Resource Description Framework (RDF) and Web Ontology Language (OWL). RDF is similar to logic programming language, focused on the facts (fact - oriented graph with two vertexes) and rules (if … then …). With the help of OWL language it is possible to describe the ontology concepts which are used in the RDF documents. Both can be considered as the semantic document's creation tools including knowledge description (OWL) and knowledge processing rules (RDF).

RDF approach for describing of information can be characterized as decomposition or splitting of information into smaller parts. There are used three-seat cortege (subject – predicate – object). The subject is entity, which is described and is located in the beginning of cortege. At the end of the cortege is an object, which describes the subject. Predicate describes the link between subject and object and is stored in the middle of tuple. Combining the individual tuples, it is possible to create a single common information presentation in the graph form.

OWL performs conceptualization of the specific domain knowledge. Created ontology consists of classes, properties and instances. Classes are formed in hierarchical structures. The common use of OWL and RDF allows making the logical deduction. Both in DBS included rules and user defined rules can be used.

4. The research tasks and required methods for fruit trees biological resources

Within the framework of the European Social Fund project "Consolidating research capacity in fruit gardening and forestry by using IT support in search for environment-friendly cultivation and product development methods", drawn up by the Latvian State Institute of Fruit-Growing, the Latvian State Forestry Research Institute "Silava" and the Riga Technical University, the research IS for biological resources of dicotyledonous plants was designed and carried out.

In tree-growing research significant is the issue how to increase the productivity of fruit varieties and quality of the fruit. It is important to make the research on determinants, which influence productivity, quality and

sort research and selection. For such tasks multidimensional analyze need.
A further direction where the fruit farmers' research problems are dealt is the research of the pathogenic organisms and development of new methods for struggle against diseases. The pathogenic population structure is changing together with the changes of growing technology, sorts and climatic conditions. Effective analyze method in this situation is multicriterial optimization.

There is also important the knowledge about plant/ sort resistance mechanisms, as well as about the characteristics of the products and their packaging. Such complex research involves a huge and diverse information and data, which are intrinsically linked and down streamed, thereby the processing and systematization needs information technology's specialist support. To analyze this process rules and deduction mechanism must be used.

5. Common architecture of the research information system

Research IS is carrying out many and various tasks. It is a dynamic system, were often it is necessary to create enhancements and improvements. Therefore, it should be unified and flexible in relation to the structural changes. As the basis for unification "information processor" models can be performed. Two alternatives were analysed.

In the first version, the research information system was made by the "information processors" network, each of which has its own specialization (see Figure 3.). "Information processor" was made as a unified body, which includes the mathematical calculation, data transformation, data visualization software packages, database systems, spreadsheets and text processors. The frame allows to connect already existing programs to "information processor", as well as to include them into the total information processing flow.

The research IS, which is created by using the "information processors" network architecture, from the beginning already focused on problem solving of data import and export. The opportunity to adapt new challenges and requirements are good. "Information processors" data input and output subsystems are oriented on different data formats and definition languages for executable functions, using templates technology. Of course, in such de-centralized system the issue of coordination and executable function's effective dividing is still topical. Problematic is also the issue of creating the unified knowledge system.

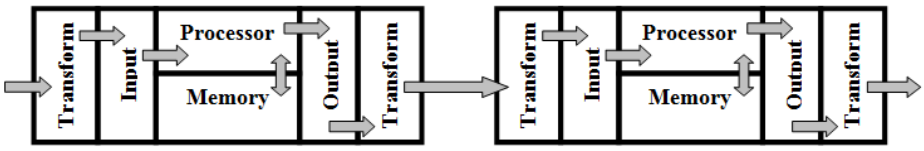

Figure 3. Research IS with "information processors" net architecture.

The initial stages of design and development pointed at another negative factor of using "information processors" network - large time and labour-consuming nature. Therefore, the overall architecture of the system was modified, focusing on "in database processing" technology and one large improved "information processor"(see Figure 4.). It was possible due to:
1) different data type storage in DB: scalar data in a relational DB, complex objects in the object-relation DB and XML data in XML DB. Using the object-relation DBS other complex data types (specific to fruit trees biological resource data) were defined, stored and also used.
2) existence of transformations in DBS from one data type to another, thus allowing implementing a virtual database, as well as making possible to retrieve data from DB in required data format.
3) unified storage of data processing methods in the DB. Using the stored procedure technology, it is easy to create the new stored procedures in different programming languages (PL/SQL, Java and other).

4) use of a single communication language in the SQL, for query execution, and in the processing of data;
5) introduction of information processing models in the DB;
6) in DB included semantic technologies, which allows establishing the knowledge base. The knowledge base was created as a deductive DB. The rules were kept in data storage structures (there was analysed the alternative form of using stored procedures) and the DBS inference mechanism was used.

According to the extended model of „information processor", two more technologies in the research IS were included. To simulate the use of the sensory memory, the external file and program link mechanism was practiced. In its turn, the short-term memory was modelled along with the mechanism of the involved systems. In this case, there is used the best understandable and the most convenient for humans' universal data processing software package – spreadsheet. It is possible to request the necessary data from DB using SQL language queries, as well as to perform data browsing and realize small calculations, and final results easily can be written again in the DB. You can also read an external document file and write the necessary information in DB.

De facto data exchange standard is XML language. Therefore it is very essentially to provide XML data type storage in DB (use of XMLType objects). Data should be loaded in XML format and then the data can be restructured by automatic data type convertors in the virtual DB.

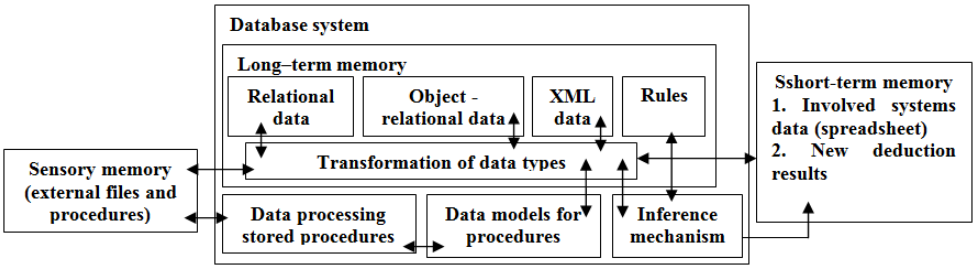

Figure 4. Realization of expanded "information processor" model.

6. The implementation of main data processing subsystems in the IS for fruit tree biological resource research

The use of extended technologies for „information processor" model, „in database processing" and „in database analytics" enables to realize the research IS easy and efficiently, which includes composite and complicated methods of calculation. For studying fruit tree biological resources such methods are used in the virtual multidimensional DB: use of deductive DB and multicriterial optimization adaptive method.

6.1. Creation and use of virtual multidimensional data base

The results of the experimental data and calculation typically are stored in a relational or object - relationship tables. In order to use the multidimensional DB form, there can be created a virtual multidimensional DB from the existing tables, including „stars" or „snowflakes" data structures. Such model can be disposed by the researcher, using the research IS base communication language - SQL. There are several ways to do this:

1. Using the commands SELECT... MODEL ... PARTITIONS by … DIMENSIONS … MEASURES … RULES … With this command the virtual multidimensional data structures or the data cubes can be defined (section PARTITIOND...). You can specify the dimensions for each cube (section DIMENSIONS by...). The internal data from the cube, which can be aggregated for obtaining the overall results (section MEASURES …). This is a unique non procedural language SQL command, because the first time in the SQL there is included the procedural instruction (section RULES ...) for defining the rules. Rules can be used for data forecasts and dynamic updates.

2. Using the SELECT command GROUP by CUBE construction to define multidimensional structures of data aggregate functions values For defining multidimensional structure data aggregate functions and the values selection there are used such functions as GROUPING() and GROUPING SET().
3. Using the SQL command DIMENSION to define expanded dimensions definitions for virtual multidimensional structure.

Virtual multidimensional DB extensive use is built on:
1) unified IS virtual DB use. All data are available in the required format (data type automatic transformation);
2) one universal language (SQL) for the virtual multidimensional DB creation and query definition. In addition, the research IS also includes the editor, a visual form for SQL command input.

6.2. Subsystem of deductive data base implementation

In the data processing many clarified causal relationship should be taken into account. Therefore, in DB the rules have to be stored. Such rules can be used together with data for deductive inferences. For example, defining allele's heredity rules, using allele's genetic data, it is possible to find the relationship among the samples in the form of potential parent – descendant combinations. The DB content for this task is following:
1) facts (predicates) stored in the deductive database: **sample(allele 1, allele 2)**;
2) rules (predicates) stored in the deductive database:
   a) **parent**(sample(X, Y), sample(X, Z)) or (sample(X, Y), sample(Y, Z)),
   where sample(X, Y) – parent sample; sample(X, Z) or sample(Y, Z) descendant sample,
   X, Y, Z – alleles;
   b) **descendant**(sample(X, Y), sample(X, Z)) or (sample(X, Y), sample(Y, Z)),
   where sample(X, Y) – descendant sample; sample(X, Z) or sample(Y, Z) parent sample,
   X, Y, Z – alleles.

Inference mechanism in the research IS obtains the data, which initially does not exist in the DB. For deductive DB implementation new „database semantic technology" was used. This technology turned out much easier to realize in comparison with classical static connection, dynamic connection or integrated CPR (Coupling Prolog to Relationship) systems.

6.3. Mulicriterial optimization realization subsystem

The main tasks of the research work are detection of causal relationship and acquisition of the best and most efficient variant. The typical mathematical model of the second group tasks is the task of multicriterial optimization. Quality criterion (global criterion) is the vector Q (X) = Q (x1, x2, ..., x_n), which consists of individual quality indicators (local criteria) q_1 (X), q_2 (X), ..., q_k (X). Looking for the solution X**or solutions set whose elements are better for all local criteria, formal multicriterial optimization solution is the Pareto set solutions.

It is necessary to determine which Pareto solutions most of all satisfies the researchers requirements. To solve this problem an adaptive multicriterial optimization approach is used. From Pareto set starting solution is selected $Q(X_0)$. The researcher defines his requirements (information) I_1 for the improvement of the criteria values. On its base the next Pareto set solution has been found. Iteration process continues until the researcher (decision making person) has convinced that better (subjective) solution in Pareto set does not exist:

$$Q(q_1(X_0), q_2(X_0), \ldots, q_k(X_0)) \xrightarrow{I_1} Q(q_1(X_1), q_2(X_1), \ldots, q_k(X_1)) \xrightarrow{I_2} Q(q_1(X_2), q_2(X_2), \ldots, q_k(X_2)) \xrightarrow{I_3} \ldots \rightarrow X^{**}$$

The solution of the multicriterial optimization task includes several subtasks: obtaining quality criteria models from experimental data; search for Pareto solutions; modelling decision maker's priorities system adaptation

of the parameters of the optimization algorithm according to the specific task (see Figure 5.). For the successful realization there is used an adaptive multicriterial optimization approach and „in database analytic" technology. In this case the data processing methods are implemented as DB stored procedures. Additionally the models of the methods are created for the values storage of indications and parameters. It allows the researcher to make the multicriterial optimization task solving flexibly (a lot of changes, concretizations).

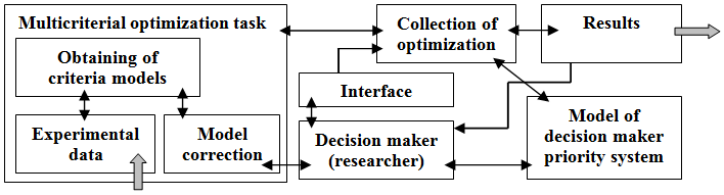

Figure 5. Solution process of multicriterial optimization.

7. Conclusion

The use of the extended model of „Information Processor" for the research IS in common architecture design confirmed the effectiveness of this variant. Very relevant to the implementation of this model are „in database processing" and „in database analytics" technologies. The complex research methods unsophisticatedly were included in the system.

According to the users' views, the developed fruit trees biological resources research IS is simple for understanding and it is easy to perform complex and difficult research tasks in it as well as there is no problem to include the innovations. Such research IS are the first research IS in Latvia. The authors also have not found the analogues in other countries.

The efficiency and flexibility of produced research IS could be increased by using the stored procedures instead of the technology of database cartridges.

References

[1] Kishor Vaidya. *Inter-Organizational Information Systems and Business Management: Theories for Researchers.* IGI Global, 2011.

[2] A. S. Nakagava. *LIMS: Implementation and Management.* Royal Society of Chemistry, Cambridge, Thomas Graham House, 1994.

[3] "How diferences in technology affect LIMS functionality, cost & ROI. System architecture strengths and limitations". LABVANTAGE Solutions, 2011.

[4] *"Research process and how a CRIS can support it*". The European Organization for International Research Information. http://www.eurocris.org Last accessed 15. October 2011.

[5] Alberto Labarga. "*Scientific Data Management. Some thinkings on scientific data management systems, LIMS*". etc. http://www.slideshare.net/alabarga/scientific-data-management Last access 11. February 2013.

[6] T. Anderson, H. Kanuka. *e-Research. Methods, Strategies and Issues.* CAAP, 2010.

[7] Newell A. *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press, 1990.

[8] Olson, J. R., Olson, G. M, "The growth of cognitive modeling in human- computer interaction since GOMS". Human-Computer Interaction 5, 1990, 221–265.

[9] V. E. Ferragine, J. H. Doorn, L.C. Rivero. *Innovations in Database Technologies and Applications. Current and Future Trends*. Information Science Reference, New York, 2009.

[10] Bill Schmarzo. *Big Data. Understanding How Data Powers Big Business*. Wiley, New York, 2013.

Authors

Principal Author: Jānis Eiduks Dr.sci.ing., asoc. prof. in Riga Technical university, System theory and design department.. Specialization: database and information system design, multicriterial optimization.

Co-author: Ainārs Auziņš holds M.sc.ing degree in Electrical mechanics and M.sc.ing degree in Computer systems. Lines of action: document management systems and web development, large databases, data warehouse and data analysis. At present he is a system analyst in Riga technical University.

Co-author: Gunārs Lācis holds PhD degree in Biological Sciences - Molecular Biology. Research activities: Genetic research of the fruit plant, the use of molecular genetics research fruit plant genetic resources. At present he is a leading scientist at the Latvia State Institute of Fruit-Growing.

Co-author: Inga Moročko-Bičevska holds PhD degree in Biological Sciences. Research activities: research on pathogenic fungi and bacteria of fruit crops, diagnostics of pathogenic fungi, establishment of testing, maintenance and growing technologies for nuclear stock and pre-base planting material of fruit crops. At present she is a leading scientist at the Latvia State Institute of Fruit-Growing.