DESIGN OF A DATA ACQUISITION SYSTEM FOR BIOTECHNOLOGICAL SYSTEMS

— research paper —

ION DAN MIRONESCU*[1], MONICA MIRONESCU**, MAYA IGNATOVA***

** Department of Chemistry and Food Engineering, Faculty of Agricultural Sciences, Food Industry and Environmental Protection, "Lucian Blaga" University of Sibiu, Romania*
*** Department of Food Biotechnology, Faculty of Agricultural Sciences, Food Industry and Environmental Protection, "Lucian Blaga" University of Sibiu, Romania*
**** Bulgarian Academy of Sciences, Institute of Control and System Research, Bulgaria, Sofia*

Abstract: The paper presents the development of a data aquisition system for the optimal storage and use of information generated by a research project. The architecture of the system is described. The design choices and the implementation steps are discussed with respect to the intended system functionality. The capabilities of the developed system as revealed by the preliminary tests are summarised in the conclusions.

Keywords: data acquisition system, semantic Web, biotechnology

INTRODUCTION

The interest to investigate the correlations between biopolymers structure and functional properties is increasing at the international level. Such correlations are used, for example, to describe the relations between the microstructure and the phase transformations (Marques et al., 2002) or to explain the structure formation for polysaccharides (Martinez el al., 2004). The analysis of colloids microstructure, devloped at the end of 90', showed that the microstructure can be characterised quantitatively by using physical, optical and rhelogical properties of materials (Quevedo et al., 2002).

[1] Corresponding author. Mailing address: University "Lucian Blaga" of Sibiu, Faculty of Agricultural Sciences, Food Industry and Environmental Protection, Str. I. Rațiu 7-9, 550012 Sibiu, Romania. Phone: 0040/269/211338. Fax: 0040269212558. E-mail address: ion.mironescu@ulbsibiu.ro

Model building is a superior modality to synthesize and represent the information aquired trough this investigation. A model reduce the incertitude and make the search for quantitative laws easier by highlighting the general paterns or the relations existent in nature (Dunn et all, 2003) (Patwardhan and Srivastava, 2004).
The huge amount of experimental data generated by the research for new product and/or technologies can be used to build such models, but requires proper storage and handling. This can be achieved only by adding data, metadata and content management to the data acquisition system (Mironescu, 2005). Because the results are obtained at multiple working places and they are concurrently needed by different users or teams, the literature recommends the use of a distributed and collaborative system that can be integrated with the (grid) computing system used for modelling and simulation (Rajasekar et all, 2001)(Rajasekar et all, 2003)(M. Valle et all, 2005). This integrated data acquisition system form the basis for a future knowledge management system.
This paper presents the development of a data acquisition system used for collection, analysis, ordering and storage of informations generated by a research project on a microbial polysaccharides: bioproduction, polysaccharide characterisation and build of the structural and functional model.

TARGET PRODUCT AND BIOPROCESS

The data aquisition system build in this work will be used for a novel biopolymer, a polysaccharide synthetised by an extreme halophilic microorganism, the archaeon *Haloferax mediterranei*. The polysaccharide has the main characteristics: contains sulphate ions and aminosugars (Anton et all, 1988) (Mironescu, 2006), which give him antiviral and antimicrobian action (Rodriguez-Valera, 1995); has high viscosity and is resistant to high salinity, temperatures and pH (Boan et al., 1998) (Mironescu, 2006); has a strong ampholitic character, being a good ions binding agent, especially for the calcium ions (Mironescu and Mironescu, 2004); is capable to form biofilms (Mironescu and Mironescu, 2006). Because of the high biotechnological potential of the microoganism and of the possibility to identify some specific functional properties of the polysaccharide, the research on the bioprocess and on the biopolymer characterisation is considered interesting and justified.
This bioproduct can be considered as a model polysaccharide for the study of other polysaccharides metabolised by extremophilic microorganisms.

This work is part of a research project structured on three levels:

- Biotechnology: obtaining of a polysaccharide with a novel microorganism and analysis of the chemical and physical factors influencing this bioprocess. The cultivation plant has the design presented in (Mironescu and Mironescu, 2009) (Mironescu et all., 2007) and consists on a lab-scale bioreactor with accesssories.
- Polysaccharide characterisation: composition, structure, conformation, molecular mass and functional properties, behaviour in the presence of other molecules.
- Construction of the structural and functional model. The model has the purpose of making correlations between structure and functional properties, and allowing the building of microstructural models of this polysaccharide individual and/or in interraction with other molecules. Finally, the model will serve as basis for a decision support system in the design and control of processes for polysaccharidic structures with improved properties.

DESIGN AND IMPLEMENTATION OF THE DATA AQUISITION SYSTEM

To support al of the levels of the project a three tier arhitecture presented in figure 1 was developed. The components, their intende functionality and their design and implementation are detalied below.

Implementation of the primary data storage (back-end)

The requirements for the system are presented.
The knowledge management system should provide:

- Distributed acquisition of data resulted from the scientific research. Because the data necessary to formulate and to adapt the model will be obtained as result of measurements and experiments in different points and by different persons, the system has to assure the consistent data centralisation. This implies a central data repository wich allows concurent access for software clients distributed in the data acquisition points. The software clients will collect directly the experimental results or will allow the entering of experimental data by the human operator.

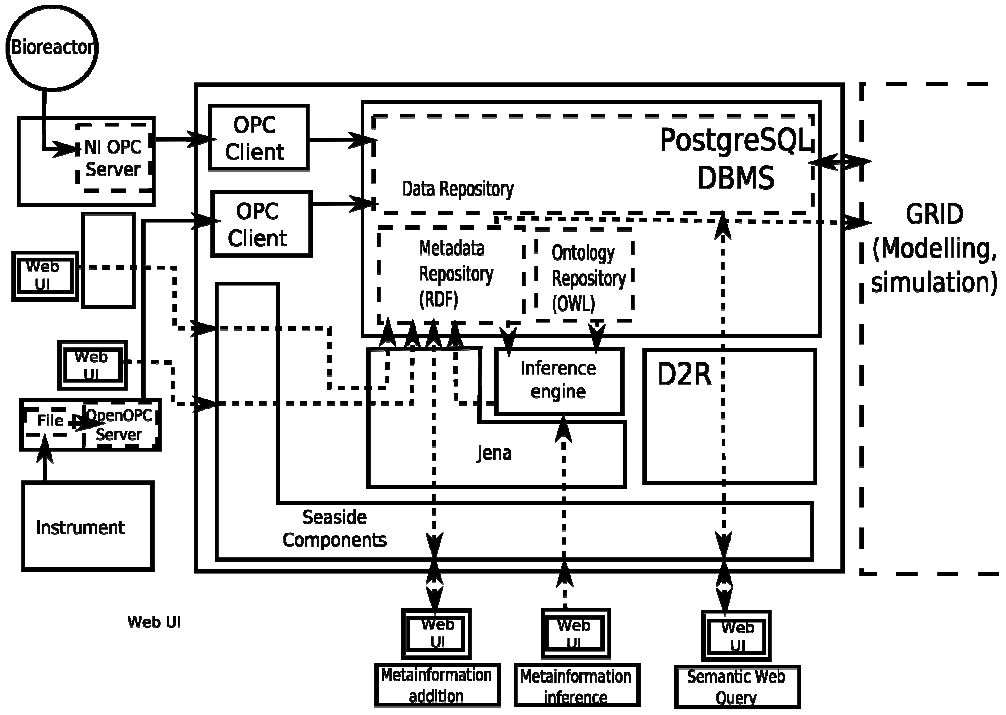

Figure 1. Architecture of the data

- Structured data storage, the database scheme will support three categories of views:
  - *Acquisition views*, for the insertion of experimentally data in the database tables. They group the data as function of experiment, acquisition point and the measured value in chronological order.
  - *Modelling views,* for the model building through:
    - Statistical filtering to eliminate irrelevant data;
    - Machine learning unsupervised and supervised (clustering, neuronal networks).

    They group the data after the causal relation input-output as sets of learning data and responses obtained from the system.
  - *Project management views*, which allow the analysis of project steps. They group the data after the person who generated the data, the experiment and the project step.
- Knowledge generation on polysaccharide structure and functional properties. The system has to offer support for adding and handling the

metainformation that allow the organisation of experimental data in knowledge on polysaccharide which is directly incorporated into the model.

- Distributed and safe access to the system data and system knowledge. The system has to:
  - allow the concurrent and distributed access of users with access rights to experimental data and knowledge;
  - ensure the data consistency and integrity.

  The system must have mechanisms for authentication and authorisation, for data access rights specification and for data protection through replication.

Based on these requirements, the offer of free and open Data Base Management Systems (DBMS) was analysed. Taking into account the requirements presented before, the selection was reduced to two products able to satisfy the entire requirements specific to this project: PostgreSQL and MySQL. For this work, PostgreSQL was finally chosen; this is an objectual relational DBMS with BSD type license. Whereas MySQL favours query speed, high degree of concurrency and easy installation and use, PostgreSQL ensure a better data integrity, transactional integrity, flexibility and enhanced extensibility, being better adapted to the project goals.
PostgreSQL was chosen also because it provides:

- support for different programming languages at server level (stored procedures) and at client level (interfaces libraries); this simplifies integration with existing acquisition infrastructure
- the capacity to rewrite by rules the operations on views allowing their flexible adaptation to the necessities;
- an interface to R language through PL/R; statistical filtering and analysis can be implemented directly in executable procedures in DBMS;
- good connexion with the developing framework RDF Sesame, which allows metadata description and handling;
- an authentication and authorisation system that can be integrated with the current authentication system, so the users administration is simplified.

The operating system used was Linux Debian 5.0. for AMD64 architecture. For DBMS, the Postgres Plus Standard Server distribution from Enterprise DB for Linux x86-64 was used on basis of the supplementary modules and the supplementary integration and testing.

implementation OF The intermediary layer

The intermediary layer has to fulfil following tasks:

- data collection from the acquisition hardware of the measurement instruments and devices used in experiments
- data insertion and update in the database;
- metadata and meta-information generation and insertion, together with data, necessary for the management of the data related knowledge and for the generation of new metadata.

To fulfil these requirements, the characteristics of the acquisition system were analysed. Data will come from two types of sources, which differ by: generation rate, duration of acquisition session, data access interface and intermediary storage way. The data obtained from the bioreactor acquisition system are generated in sessions that can last days, with sampling in the order of minutes or even seconds and with 7 to 10 parameters measured at the same time. For a correct data sampling, the acquisition system is running its own OPC server, which assures the intermediary data storage. The server is accessible over Ethernet, so that distributed access is assured.

Data resulted from devices used to measure some characteristics are generated in session that lasts hours, with sampling in the order of seconds or less and with 1 to 3 parameters measured at the same time. Values are collected by software specific to each device and which uses local protocols (RS232 or USB) to communicate with the acquisition hardware. Data are saved at the end of each session in text or binary files.

Based on this analysis, the working environment for the intermediary layer was chosen. So, for the data generated by the experiments in bioreactor, a development platform with OPC support has to be chosen. After comparing the existent solutions, the Python language was chosen because it offers:

- open/free toolkit for creating OPC client or server applications;
- support for visualisation and processing of scientific data;
- support for connecting with PostgreSQL;
- portability on Windows and Linux;
- capacity to “glue” pieces written in static languages (C/C++).

In order to create a unitary interface, applications in Python were developed to collect the data from each experiment and to expose them as web services through OPC XML DA.

For metadata description a semantic web platform with RDF (Resource Description Framework) support was used.

Because the data will be stored in a classical database, a mapping between ontology and relational scheme has to be chosen. The D2RQ platform was selected for this task. The platform includes a language for mapping description and a D2R server for access to the relational data with the instruments specific to the RDF triplets interrogation. For the ontology

description, triplets publication, interrogation and inference, the Jena platform was selected. This platform allows the inference on the triplets by using internal or external inference engines. Consequently, the semantic web part will be developed preponderantly in Java (the language of D2R and Jena). The integration between the applications written in the two languages will be assured by means of Jython and RDFAlchemy packages.
The heterogenic and distributed structure of the intermediary layer and the web-type architecture impose the implementation of the interface between the intermediary layer and the representation layer trough web services. The support offered by both languages for the web services development is consistent, simplifying the writing of service providing modules.
The implementation of the intermediary layer was realised by installing, configuring and testing on the server of:

- the virtual machines, standard libraries and the development environments for the two used languages: Java and Phython. In both cases, the solutions for the platform AMD 64 were used, for over passing the limitations of the architectures on 32 bits.
- the Java and Phython drivers for PostgreSQL database access;
- the D2R server (which includes Jena);
- the libraries necessary for developing applications on the intermediary layer.

D2R and Jena were configured to use PostgreSQL as repository and the necessary databases were created.
Prototypes were developed for the following modules:

- OPC client which accesses data from the bioreactor acquisition system;
- OPC server which reads data resulted from the measurements of the polysaccharide properties;
- Metadata access interface.

The prototypes contain the code necessary for their functioning as web services. They will be completed and configured at the beginning of data acquisition when the exact configuration and workflow for data acquisition will be established.

Implementation of the web interfaces for data and knowledge input and access (Front END)

In the context of the architecture defined before, the user interface should:

- collect the input from the user users;

- transform those inputs in requests for the services implemented in the intermediary layer;
- transmit the request to the responsible service;
- receive the answers and present them to the users.

The Seaside environment was chosen for the development of the interface layer, because it has advantages for both the application architecture and the development.

A Seaside application is built from reusable components which can interact and can be combined in different ways. Because each component has its own control flux, independent of those of other components, the implementation of an application that behaves similar to a desktop application is simplified in comparison to a classic web development framework based on pages.

Because the development environment is based on Smalltalk, the application can be debugged on-the-fly and the execution is restarted from the point where the processing was interrupted by the error without recompilation. Those characteristics reduce the complexity and shorten the development cycle for the user interface.

Two categories of interfaces were developed: administrative interfaces and user interfaces.

Administrative interfaces assure the administration of the system resources: users, access rights, registered services, data acquisition points, ontology.

User interfaces allow:

- data input by sources selection and time establishment for other acquisition parameters;
- metadata generation by inference and their addition to the data;
- interrogation of existing data and metadata.

The components that allow user interaction and the components which access the services on the intermediary layer were implemented for each type of interface.

CONCLUSIONS

The preliminary tests had shown that the implemented data aquisition system provides support for:

- distributed acquisition of the scientific data generated by the researchers; tests performed with the MatrikonOPC Simulation Server have shown that the system can support the proiected data acquisition rate of both the biopreactor and analysis instrumentation;

- structured data storage; on basis of the database schema the integration of the data storage with a scientific visulisation system was implemented in ½ hour and 85 lines of code;
- support for the generation and storage of knowledge on the polysaccharide structure and functional properties; the test data stored in the acquisition test where accesible to semantic web queryes after defining a ontology related to biopolymer characteristics;
- distributed and secure access to the stored data and knowledge;
- assisted generation of the model and the simulation case files; the script that generates on basis of the defined aquired data and test ontology the coresponding simulation case whas implemented in ¼ hour and 32 lines of code.

This support will optimise the data aquisition at the biotechnological and acquisition level and and will shorten the structure cration time, allowing for an efficient comparative study at the modelling level.

This caracteristic are consistent with the the goal of the presented work: the design and implementation of a reduced costs. dynamic, perfectible and adaptable system for data acquisition.

Acknowledgments

Financial supports from CNCSIS Romania (Consiliul National al Cercetarii Stiintifice din Invatamantul Superior) by the research grant ID_473 (2009-2011) are gratefully acknowledged.

REFERENCES

1. Anton, J,, Inmaculada Meseguer and F. Rodriguez-Valera, 1988, Production of an extracellular polysaccharide by *Haloferax mediterranei*, *Applied and Environmental Microbiology*, 54 (10), p. 2381-2386
2. Boan, I.F., Garcia-Quesada, J.C., Anton, J., Rodríguez-Valera, F., Marcilla, A., 1998, Flow properties of haloarchaeal polysaccharides in aqueous solutions, *Polymer*, 39, p. 6945-6950
3. Dunn, I.J., Heinzle, E. Ingham, J. Prenosil, J. E., 2003, *Biological Reaction Engineering Dynamic Modelling Fundamentals with Simulation Examples*, Wiley-VCH, Weinheim
4. Marques, E., Dias, R., Miguel, M., 2002, Association in polyelectrolyte-catationic vesicle systems: from phase behavior to microstructure, i in *Polymer gels and networks*, ed. by Y. Osada, A.R. Khoklov, Marcel dekker Inc., Basel, Switzerland, p. 67-101
5. Martinez, L., Agnely, F., Bettini, R., Besnard, M., Colombo, P., Couarraze, G., 2004, Preparation and characterization of chitosan based micro networks:

Transposition to a prilling process, *Journal of Applied Polymer Science*, 93 (6), p. 2550-2558
6. Mironescu I.D., 2005, Web Based solution for scientific data management, *Proceedings of the international conference Agricultural and Food Sciences Processes and Tehnologies*, p. 238-244
7. Mironescu M., Mironescu V, New concept for the obtention of biopolymers-based food biofilms, *Journal of agroalimentary processes and technologies*, 2006, XII, no. 1, p. 219-216
8. Mironescu, M., Mironescu, V., 2004, The complexant properties of exopolysaccharides produced by the archaebacterium *Haloferax mediterranei*, *Scientifical researches. Agroalimentary Processsses and Technologies*, vol. X, no.1, p. 78-85
9. Mironescu, M., Posten, C., Kriger, C., Tzoneva, R., Control and command of biotechnological processes using a flexible application developed in LabView, 2007, 4th IFAC MCPL, p. 801-806
10. Mironescu, M., *Studii şi cercetări privind producerea şi caracterizarea polizaharidelor de origine microbiană folosind microorganismul Haloferax mediterranei*, Teza de doctorat, Universitatea Lucian Blaga din Sibiu, feb 2006
11. Patwardhan, P.R., Srivastava, A.K., 2004, Model-based fed-batch cultivation of *R. eutropha* for enhanced biopolymer production, *Biochemical Engineering Journal*, 20, p. 21-28
12. Quevedo, R., Carlos, L-G., Aguilera, L.M., Cadoche, L., 2002, Description of food surfaces and microstructural changes using fractal image texture analysis, Journal of Food Engineering, 53, p. 361-371
13. Rajasekar A., Moore R., Data and Metadata Collections for Scientific Applications, 2001, *European High Performance Computing conference*, Amsterdam, Holland, June 26, http://www.sdsc.edu/srb/Pubs/Data-management_moore.pdf
14. Rajasekar A., Wan M., Moore R., Kremenek G., Guptill T., Data Grids, Collections and Grid Bricks, 2003, *20th IEEE/ 11th NASA Goddard Conference on Mass Storage Systems & Technologies (MSST2003)* San Diego, California, April 7-10, http://www.sdsc.edu/srb/Pubs/bricksMS2003.pdf
15. Rodriguez-Valera, F., 1995, Cultivation of halophilic Archaea, in *Archaea, a laboratory manual*, ed. by Robb, F.T., Place, A.R., Sowers, K.R., Schreier, H.J., DasSharma, S., Fleischmann, E.M. , Cold Spring Harbor Laboratory Press, p.13-16
16. Valle, M., Favre, J., Parkinson, E., Perrig, A., Farhat,M., 2005, Scientific Data Management for Visualization Implementation Experience, *Simulation and Visualization 2005*, Magdeburg, March 3 – 4