Long Lived Transaction Processing for Business Processes and Scientific Workflows

João E. Ferreira[1], Luciano V. Araújo[2], Kelly R. Braghetto[1],
Isabel C. Italiano[2], Márcio K. Oikawa[3], Pedro L. Takecian[1]

[1] Institute of Mathematics and Statistics
University of São Paulo
{jef,kellyrb,plt}@ime.usp.br
[2] School of Arts, Science and Humanities
University of São Paulo
{lvaraujo,isabel.italiano}@usp.br
[3] Center of Mathematics, Computing and Cognition
Federal University of ABC
marcio.oikawa@ufabc.edu.br

Abstract.

Process-oriented systems such as scientific workflows and business processes can be designed and implemented based on advanced transaction models or other formal models for concurrent distributed systems. Advanced transactional modeling approaches trade the classic benefits of isolation and atomicity properties for long lived transactions, giving it the ability of being applicable in systems that run in heterogeneous, autonomous, and distributed computing environments. At same time, process-oriented modeling based on formalisms such as Petri nets, process algebras, and graph-based process models has improved the correctness of designed processes. IME-USP DATA group has been working in important process-oriented and data-oriented challenges which typically have requirements concerning the quality of service and correctness. Among our current research topics, we can highlight: exception handling for long transactions; automated generation of business step dependency representation; data extraction, transformation, and loading for data warehouse systems; automation of clinical and molecular data processing; and formal modeling of business processes aiming qualitative and quantitative analyses.

Categories and Subject Descriptors: H.2.4 Information Systems [**Database Management**]: Systems—*Transaction Processing*

Keywords: transaction processing, scientific workflows, business process management

1. HISTORY OF DATA GROUP

The research group of *Database Modeling, Transactions, and data Analysis* (DATA Group) [1] was born at 2000, supported by *Institute of Mathematics and Statistics* (IME) - *University of São Paulo* (USP) and *Bioinformatics Center* (BIOINFO-USP). Nowadays, DATA group develops researches and several projects for academic and industrial purposes, specially in database integration; modeling, analysis, and implementation of business processes and scientific workflows; data warehouses; and asynchronous transactions. DATA group also supports undergraduate and graduate database system courses, providing technical support, seminars, lectures, and guidance in recent topics on Database

[1] http://www.data.ime.usp.br

This work has been supported by FAPESP (Research Support Foundation of São Paulo State) number 2010/15493-4. Additional support is provided by CNPq (Brazilian National Research Council) grant number 201557/2009-6.

Systems. DATA group counts upon the computing infrastructure at IME-USP and has two important available laboratories. The resources to set up these laboratories were originated especially from the following projects: CAGE [2], Malaria [3], Transaction Processing [4], Data Analysis for Blood Donation[5], and SUN-LabBd[6]. These laboratories have been used primarily by graduate and undergraduate DATA group students that develop activities linked to research projects in the field of database systems. The DATA laboratory IME-USP, room 255 block A, provides adequate infrastructure (network, air conditioning, furniture, stabilizers and no-breaks), and it is also equipped with 20 PC computers and 6 Linux servers.

The history of DATA group is related to the transaction processing [Ferreira and TrainaJr 1994; 1996; TrainaJr et al. 1997] and database system modeling [Ferreira and Busichia 1999]. The group has been originated from the Databases and Image group at ICMC/USP [7].

Inside this transactional processing issue we have been researching some topics to support a long-term solution of scientific and business process applications that require formally correct and automated processing tools for process-oriented large scale applications. Some initial results in transactional processing can be seen in [Ferreira et al. 2005a; 2005b]. Usually, these process-oriented large scale applications have been designed and implemented based on classic long transaction processing, scientific workflow engines, and Web services. However, the recovery support for long transactions still remains as a challenge to be overcome. The main contribution in transaction processing researches, which is addressed to this challenge, is an alternative approach for transactional recovery support in scientific workflows and mission critical services.

The architecture and the concept of *RiverFish Navigation Plan* [Ferreira et al. 2005a; 2005b] were our first initiative related to the transaction processing. It was used to implement a registration system of *Finance Secretary of the State of São Paulo* (SEFAZ) that receives around 2,000 requests per minute. This registration system supports working permit orders, taxes controlling, and authorization for printing of tax documents. All requests are routed through the validation of external and internal consistencies. The external consistencies validation are hold using interactions with autonomous and heterogeneous systems of other governmental institutions such as *Federal Government*, *State Company Associations*, *State Government Environment Control*, and *State Health Surveillance*.

More recently, DATA group has focused on detailed study of process algebras for support Business Processes Management (BPM) [Braghetto et al. 2007; Oikawa et al. 2009; Ferreira et al. 2006; Ferreira et al. 2010]. During this period, our main goal has been to understand the advantages and disadvantages of the process algebras and also other formal frameworks such as Petri nets. We compared these approaches with other domain-specific languages such as *Business Process Execution Language* (BPEL) and *Business Process Model and Notation* (BPMN) [Takecian et al. 2010].

DATA group also have been working with multidisciplinary projects, which involve database development and analysis. An important example of these projects is related to improvement of security in blood transfusions [Carneiro-Proietti et al. 2009; Almeida-Neto et al. 2009]. This project aims at developing a database to help blood banks and their specialists on analysis of their stored data. Two sub-projects will be developed in this domain: a) analysis of residual risk of HIV transmission; b) epidemiological study regarding the confidence of questionnaires answers obtained during blood donors

[2]Project titled "Cooperation for Analysis of Gene Expressions", supported by FAPESP, grant number 99/07390http0.
[3]Project titled "Genomics and post-genomics approaches to study the human malarial parasites *Plasmodium vivax* and *Plasmodium falciparum* in the Brazilian Amazon", supported by FAPESP, grant number 01/09401-0.
[4]"Transaction Model for Recovery Business Process Management and Scientific Workflows", supported by FAPESP, grant number 2010/15493-4.
[5]"Data Analysis of Blood Donors in Multidimensional Database", supported by CNPq (Brazilian National Research Council), grant number 201557/2009-6.
[6]"Disk Access Optimization for Very Large Database", supported by SUN Brasil.
[7]http://www.gbdi.icmc.usp.br/

screenings. Another project under development in collaboration with *Health Department of Brazil* aims at developing methods and tools to create and to analyze clinical and molecular data markers from HIV patients to help*Department of STD (Sexual Transmitted Diseases), Aids and Viral Hepatitis* on decision making [Araújo et al. 2006]. This department relies on bioinformatics methods and tools to produce its results. Therefore, this project will: 1) provide technical certification for learning the computational environment (systems DBCollHIV, HIVdag, etc.), as well as the extraction and analysis of molecular and clinical data markers; 2) apply classification techniques for molecular and clinical data analysis, as well as the automatic identification of associations between mutations and resistance to anti-retroviral drugs; 3) develop an algorithm to identify mutations from FASTA[8] files; 4) develop transactional systems to gather trustworthy molecular and clinical data available in the internet

The decision of developing research in topics like e-Science, e-Commerce, data integration and long term transactions is based on the group's mission of promoting applied research as a form of an immediate contribution not only to the academic community, but also to the whole society. This mission has allowed us to contemplate the challenges of maintenance and evolution of systems, a common feature of applied research. In this sense, the group has sought to develop research models that allow the sustainable maintenance and development of systems and enable collaboration between research groups and final users, which improves the quality of systems. Thus, the research work developed by DATA group aims the alignment with the academic pursuit of solutions for managing large volumes of data, long term transactions, scientific collaboration and automation systems. Moreover, this research work has contributed to different society concerns, like development of computational tools to support medical decision making, the study of the impact of the medication used in the treatment of patients infected by HIV, genetic research and the quest for quality of execution genetic testing, projects for security on blood transfusion, and e-government and industrial applications.

The rest of this paper is organized as follows. In Section 2 we describe our background research and our main goal. From Section 3 to Section 7 we summarize research issues and contributions according to this main goal. In Section 8 we survey our next steps, and Section 9 concludes the paper with a brief summary of our results.

2. BACKGROUND RESEARCH AND THE MAIN GOAL OF THE DATA GROUP

Large data sets of molecular and clinical data in the modern biological research institutions and business process in Web service applications have introduced new requirements for database storage and transaction processing systems. In business process Web applications, the capacity for data generation, data storage, and guaranteeing the quality of services in a large scale have been increasing in last years, specially in big international companies such as Google, SAP, and IBM. Scientific information systems have waked up for the need of robust data storage and collaborative aspects in order to face challenges and provide solutions in this area. There are two famous phrases that resume the state of art in e-Science information systems. The first phrase is "Science is becoming data-intensive and collaborative", and the second is "Researchers from numerous disciplines need to work together to attack complex problems; openly sharing data will pave the way for researchers to communicate and collaborate more effectively". These phrases were written by Ed Seidel, acting assistant director for NSF's Mathematical and Physical Sciences directorate [9]. These business and scientific scenarios show that we are in "data deluge age" where the long lived transaction processing under collaborative information system perspective is an important computer science challenge.

In typical e-Science laboratory routines, long lived transaction processing is used in many tests that are performed concurrently even by the same machine. This high parallelism demands precise

[8]Text-based format for representing either nucleotide sequences or peptide sequences.
[9]http://www.nsf.gov/news.

management of procedures, reagents and results. New tests are defined frequently, so users have to be guided to execute the right task at appropriate time. Incompatibilities among previous processes and new data requirements make the integration and analysis of available knowledge very difficult. This problem is compounded by the process of scientific knowledge discovery, which requires frequent process updates and refinement of scientific hypotheses. This scientific scenario requires a long lived transaction processing to avoid data manual approaches that quickly become very expensive or commonly infeasible.

In business process applications hundreds of billions of dollars are annually invested on business process automation, and the rapidly growing market for business process outsourcing (concretely implemented by customer and technical support call centers) underscores the importance of human assistance to supplement automated web-facing applications such as e-commerce. One of the main functions of call centers is to help customers to recover failed transactions, often resulted from interrupted sessions while interacting with automated workflow (or composite services) applications. The need for human assistance arises from the lack of automated and comprehensive failure recovery procedures, one of the major limitations of such workflow applications. These important mission-critical applications such as financial, military, and e-commerce, typically have real-time quality of service requirements (e.g. severely limited response time) that are difficult to satisfy without long lived transaction processing in order to avoid the collapse or the hard task to upgrade collaborative information systems.

Considering the business and scientific scenarios described and the challenges in data processing, the main goal of DATA group is to provide fundamentals and to develop tools to support computer applications that involve long transaction solution for scientific and business process data integration problems requiring long transaction fundamentals, performance analysis, and automated information processing. In the next sections we summarize research issues and contributions according to this main goal.

3. EXCEPTION HANDLING FOR LONG TRANSACTIONS PROCESSING

In contrast to process-oriented approaches, we have adopted the WED-flow (Work, Event, and Data flow) approach [Ferreira et al. 2010; Ferreira et al. 2010; Wu et al. 2010] to address the challenges of automated exception handling needed in scientific workflows and mission-critical services. WED-flow combines the concepts of workflow composition, transactions, events, and data states. At an abstract level, the normal path of a workflow is modeled as a SAGA[10] [Garcia-Molina and Salem 1987], composed from SAGA steps, each of which is enclosed in a transaction. SAGA steps rely on the transaction processing system to recover from internal failures. Since SAGA steps only terminate in a small number of states, e.g., either commit or abort, the number of exception cases is more manageable, proportional to the number of SAGA steps rather than the number of all possible failures.

The first advantage of WED-flow approach is the preservation of correctness properties and system consistency throughout the execution. The second advantage is the adoption of recovery algorithms and mechanisms that facilitate the design and implementation of failure recovery as exception handling routines. To explain the achievement of automated recovery through WED-flow, we briefly outline the execution of a WED-flow, which consists of a sequence of SAGA steps. These SAGA steps are connected together and triggered by events, defined as carefully written pre-conditions for each step. During the execution of SAGA steps of a WED-flow, we record all the important WED-flow execution information at each step boundary. This record (called WED-state) includes data changes such as the old and new values of data items modified by each SAGA step, and application-wide integrity constraints that need to be maintained. For both the normal path and exceptions, an event triggers the

[10]SAGA is the name of the transactional model that was suggested by Bruce Lindsay. There is no acronyms, but rather a reference to long stories about various events or people from the past.

execution of an appropriate following SAGA step when its pre-conditions become true. For the normal path, each SAGA step guarantees continued control flow at its own termination by making sure the following steps pre-conditions are satisfied. On the other hand, if an exception arises and the execution deviates from the normal path, e.g., scientific experiment cancellation or step experiment rejection, the pre-conditions will trigger either a backward recovery or a forward recovery. These recovery steps use WED-states to restore database consistency either to a previously consistent state (backward recovery or compensation) or an alternate consistent state (forward recovery). These recovery routines may be reused through careful chaining, e.g., successive backward recovery steps that return the system to the initial state of WED-flow, regardless of which step initiated the recovery.

In terms of transaction processing issues the first goal of the research described in this section is the development of a theoretical model that captures the QoS (Quality of Service) properties of a composite service, given the QoS provided by components. The second goal of this research is the development of software tools that support the application to implement realistic, but demo-scale mission-critical applications such as financial transaction support. The third goal of this research is an experimental evaluation of the effectiveness using a realistic application, e.g., the financial application mentioned above.

4. THE USE OF FORMALISMS TO SUPPORT BUSINESS PROCESS MANAGEMENT

The *Business Process Management* (BPM) comprehends the languages, methods, techniques, and computational tools created to support the business processes in all phases of their life cycle – i.e., modeling, implementation, execution, monitoring, and analysis. The *Business Process Management Systems* (BPMS) can be seen as a natural evolution of *Workflow Systems* (WFS). In WFS, the main preoccupation was the automation of the operational processes. With the improvement of the modeling techniques and automation tools, the BPMS could dedicate more attention to the quality of the processes.

Most part of the languages and methods specifically created to the modeling of workflows and business processes do not have a formal foundation. We can cite as examples of non-formal modeling languages the *Business Process Model and Notation* (BPMN), the Activity Diagrams of the *Unified Modeling Language* (UML), and the *Event-driven Process Chains* (EPC). Despite being largely used in practice, these modeling techniques are not the most appropriate ones to support the analysis phase, neither to support a reliable implementation of the modeled behavior. The lack of formal semantics for the constructors of a modeling technique may cause ambiguities in the interpretation of the models and forces the management tools to make their own assumptions about the meaning of the constructions.

At DATA group, we have addressed this problem studying the application of several formal techniques in the modeling of business processes. We started our research with two classical classes of formalisms broadly used to specify concurrent and distributed systems – the *Petri nets* and the *process algebras*. We analyzed these two classes and compared them under the BPM domain; both of them presented pros and cons, and demonstrated to be effective in the modeling of business processes. However, differently from Petri nets, the potential of process algebras in this research domain has not being well exploited yet. For this reason, the DATA group focused its studies on the use of process algebras to support the modeling and the execution of business processes.

One of the main results of our work in this research area was the creation of the *Navigation Plan Definition Language* (NPDL) [Braghetto 2006; Ferreira et al. 2006]. The NPDL is a language for the specification of business processes based on the *Algebra of Communicating Processes* (ACP), a very powerful yet simple process algebra. In a process algebra model, the behavior of a system is defined as algebraic expressions. These expressions are formed by atomic actions composed by operators that indicate the execution order of the actions. NPDL implements the most important operators of

ACP and also defines additional operators that help us to model more complex process behaviors. With NPDL, we are able to model the most frequent control-flow patterns, as shown in the work of Braghetto et al. [Braghetto et al. 2007].

To support NPDL, we implemented the *NavigationPlanTool* (NPTool) [Braghetto et al. 2008a; 2008b; 2009], a software tool to control the execution of business processes that can be easily integrated into the information systems. NPTool implements the NPDL as an extension of the SQL language and offers two other important services: processes instantiation and process instances execution monitoring. With NPTool, we showed that process algebra features could be combined with a relational database model to provide a scalable and reliable execution control of business processes. The NPTool has a web service interface – the *Navigation Plan for Web Services* (NPWS) [Rodrigues et al. 2009] – which enables its use in the composition of web services. The NPTool and the NPWS are being used by several deployed systems developed in our research group, such as the CEGH System (Section 7.1), OCI 2 -Optimal Clones Identifier by Dynamics Rules [Cantão et al. 2010] and the HIVDag [Araújo et al. 2008].

Another important result of the DATA group related to the formal modeling of business processes was the work of Takecian et al. [Takecian 2008; Takecian et al. 2010]. This work evaluated the viability of using the *Language of Temporal Ordering Specification* (LOTOS) to model business processes and workflows. LOTOS is a very rich extension of process algebra, used to model both the dynamic behaviors of systems, data structures and value expressions. The work also compared the expressiveness of ACP and LOTOS, using as comparison basis the workflow patterns (i.e., the most common routing structures found in workflow models) and *Business Process Execution Language* (BPEL) concepts.

All the results mentioned above are related to the *quality* of the model and, consequently, the reliable execution of business processes. However, more than help to analyze and improve the quality of the processes, some formalisms enable the *quantitative analysis* of the processes. The most important type of quantitative analysis is the *performance evaluation*.

There are three approaches to evaluate the performance of systems broadly discussed in the literature: the *analytical modeling*, the *simulation* and *measuring*. Both analytical modeling and simulation are model-based approaches, but the former is the only one that requires a formal model to be made. From the solution of an analytical model, we can extract several performance indicators, such as the utilization rate of the modeled resources, the throughput of the tasks, the service time (e.g., the average time to complete the execution of a process instance), etc. With these indicators, we can identify inefficiencies and optimize business processes.

At DATA group, we are currently studying the use of variated stochastic formalisms aiming the performance evaluation of business processes via analytical modeling. We have compared the application of three different classes of Markovian formalisms –*stochastic Petri nets*, *stochastic process algebras* and *stochastic automata networks* – in the modeling of different business process scenarios. The identified pros and cons of each formalism were described in the work of Braghetto et al [Braghetto et al. 2010].

Having in mind that the stochastic modeling of a system is not a trivial task, specially for designers with no statistical skills, we created a framework to automatically convert business processes modeled in BPMN to stochastic automata networks (SAN) [Braghetto et al. 2011]. SAN enables us to build compositional models and attenuates the state-space explosion problem associated with other Markovian formalisms; it is used to model and analyze large/complex systems. Our framework was implemented in a software tool – the BP2SAN – that is publicly available at [11]. To improve the performance analysis, we are developing new approaches to consider resource management in business process models. The tasks of a business process frequently depend on resources (e.g., humans, machines, hardware, software, etc.) to be executed. It is well-known that the resource contention is

[11] http://www.ime.usp.br/ kellyrb/bp2san.

a factor that greatly impacts the performance of a system. For this reason, it is important to understand what are the resource requirements of a business process and how they could be appropriately modeled.

5. AUTOMATED GENERATION OF BUSINESS STEP DEPENDENCY REPRESENTATION

One of the current important problems on BPM is how to create business process models with both formal correctness and flexible manipulation for end users. Formal approaches for evaluating correctness and business process models behavior are usually complex and require some expertise from their users. These languages, such as the process algebras, cover a set of grammar rules that guarantee formal correctness when well applied. Meanwhile, a significant part of business models has been being empirically built, through graphic software tools, privileging the experience and knowledge of end users, but disregarding formal validation. Graphic tools normally use graph based models and interactive processes for defining business models. Since such models are usually dynamic and evolves through time, models without formal verification tend to suffer inconsistencies after maintenance procedures.

Therefore, we can identify a clear gap between business model modeling and formal verification. Graphic tools usually does not completely guarantee correct models. This occurs because all elements inside a business process are dependent each other. For example, some execution rules, like splits and synchronizations, depend on previous rules to generate correct paths for all instances. Some wrong combinations of rules may include structural anomalies inside the process models, such as deadlocks and redundant subpaths, with no natural mapping in process algebra. This conceptual difference makes difficult to relate graphs and process algebra, specially on exceptions or defective constructions. Therefore, there is no concrete and formal conclusion about the expressiveness of graph models compared to process algebras, what does not offer safe limits to the application of mapping techniques.

One of the first investments of DATA group at this area was to use control-flow workflow patterns as reference to validate modeling techniques for business processes. Using Riverfish architecture, the paper [Zuliane et al. 2008] explores the application of the main patterns as a starting point for creating correct business process models. These new modeling approach supports end users on building processes using pre-defined graphic fragments.

Even though this modeling technique avoids some serious problems on many models, it was not enough to ensure the correctness of them all. So, DATA group has been studying some ways for linking graph based models and process algebra evaluation, in order to use the main advantages of both. In [Oikawa et al. 2009], the group has introduced expression graphs, which are directed acyclic graphs with algebraic elements (operators inside vertices and sub-expressions inside arcs). Expression graphs were used to guide the generation of process algebra expressions from initial graph-based models. Among the contributions of this work, we should indicate the definition of a class of graphs for which is possible to easily define a correspondent algebraic expression. In addition, a proposition of an algorithm that analyses these graphs considering two main objectives: to generate algebraic expressions for well-formed expression graphs and to identify anomalies (deadlocks and redundant subpaths) on defective expression graphs.

Even if there are only well-formed expression graphs, there are some structural forms that hamper the definition of algebraic expressions. In general, all of them are derived from Wheatstone bridges. Many process algebras present difficulties on representing Wheatstone bridges, since they have to map all valid execution path combinations. Most of the doubts about linking process algebra and graphs may be reduced after treating consistently Wheatstone bridges and their variations. The next steps of the group comprehends the definition of new techniques for mapping generalized Wheatstone bridges into algebraic expressions. Another research interest is to extend this method to directed graphs with cycles, studying their influence both on the algebraic expressions complexity and the inclusion

of anomalies into original structure.

6. DATA EXTRACTING, TRANSFORMATION, AND LOADING FOR DATA WAREHOUSE PROCESSES

Due to the fact that experiments in e-Science have a process of knowledge discovery in common with each other, they undergo frequent updates, either because of a more refined vision on the part of the scientist or because of changes in the steps and tasks that are no longer relevant to the biological procedure. The re-use of knowledge generated by experiments allows scientists to better manage resources and time. The results from an e-Science experiment will have little use if the scientists are not able to judge its adherence to the problem, to characterize how trustworthy the data is, to propose model hypotheses and, finally, to redefine or define new experiments.

Such data is stored in a large database called data warehouse. A data warehouse summarizes the data that are organized in dimensions, making them available for searches and analyses through On-Line Analytical Processing (OLAP) applications and decision support systems. Implementing a data warehouse consists basically of collecting, cleaning and storing the information that comes from a variety of sources by way of a periodic transfer of data. Once created and the data loaded for the first time, the data warehouse can receive incremental updates, which should be kept in sync with the operational base, thus transforming it into a base with a large volume of data. Our DATA group have been working in two parts in data warehouse issues.

6.1 Part 1: Practices to Reduce the Complexity of Data Warehouse Development

Recently, it has been increasingly frequent the use of Data Warehouses (DW's) to analyze, uniformize and centralize data coming from large systems. These data often come from several heterogeneous transaction systems. To build such DW's is a very challenging task, mainly due to the considerable size of the projects and to the difficulty existent in the interactions of people with different profiles. Because of these factors, many times these projects either never end or become obsolete even before they are ready for use. In the last years, the DATA group has had an experience that fits the scenario described above in a project called *Retrovirus Epidemiology Donor Study* - II (REDS-II) [Carneiro-Proietti et al. 2009; Almeida-Neto et al. 2009].

The REDS-II project is composed of a network of United States blood centers sponsored by the *National Institute of Health* (NIH) USA with the purpose of developing research projects focused on blood safety. Three important Brazilian blood centers were also included. Our main responsibilities in this project have been performing collection, cleaning, standardization, storing, analyses, and reports generation related to Brazilian blood centers data. Each blood center has its own transactional system, suitable for its local needs and requirements.

To mitigate the difficulties pointed out, during the development of REDS-II DW we have applied some well-known practices in a combination that would help us to achieve a successful implementation. This development technique was explained in a paper called "Good Practices to Reduce the Complexity of Data Warehouse Development for Transactional Blood Donation Systems", still in a journal submission process. Here, we will briefly describe some ideas involved in this work.

To reduce the complexity of a DW development in a complex scenario like blood donation, we suggest the combined use of three good practices: conceptual modeling, data analyses and a modular architecture. When we deal with different blood donation transactional systems, we have to work with different models because they reflect different processes (each blood center has its own donation process). However, fundamental concepts tend to be the same in all centers. The conceptual model helps us to find out these common basic entities, the relationships between them and assists us to build the DW structure. Once we have a working version of the DW, it is time to populate it and

do some analyses to verify the existence of errors in the concepts that have been modeled so far. After that, it is possible to develop a new version of the conceptual model, with the errors corrected, and this leads to an improved DW. Cycling the stages of analyses and corrections (in the model and implementation) we achieve a stable and functional version of the DW. Since stability is reached, we can think in expanding the conceptual model to cover more concepts, restarting the cycle of analyses and corrections. To support such developing process, full of changes and expansions, the DW system must have a flexible and loose coupled architecture. So, we suggest the use of a modular architecture, which improves the system organization, facilitates its maintenance and its growth. Our system is composed of four modules. The first one is responsible for receiving and temporarily store data from blood center's transactional systems in a stage area. The second module is composed of data treatment and homogenization routines. It is responsible for cleaning the data coming from the first module and transforming them so as to neutralize differences arising from heterogeneity of data sources. Module three integrates all data into a normalized and historical database. This database will permanently store data from all blood centers. The last and fourth module is focused on data visualization and analysis. Here, we build analytical databases to facilitate the generation of cube views and the processing of more elaborated queries. By using cubes, non-expert users in computer science or engineering can discover relationships between different entities attributes using software tools they are habituated to.

Until now, in REDS-II project, the combined use of these practices has proved to be highly efficient and has brought excellent results, showing itself to be very convenient to develop such kind of system.

6.2 Part 2: Synchronization Options for Data Warehouse Designs

In order to delivery data in real time, many projects are integrating just-in-time data and the updating between transactional and analytical databases has become a critical issue. Some areas require increasingly faster and more accurate decision processes. To achieve maximum efficiency, these business models require feeding transaction data to a data warehouse in real time or at periods shorter than traditional static loadings. Most data warehouse designs view the entire data warehouse as a single and homogeneous set of information in terms of synchronization. However, data warehouse portions represent different business models and have their own update requirements that change over time as data warehouse evolves. It is important to keep track of the data warehouse changes in order to adjust its synchronization model and achieve the most suitable option for data loading periods. To address this problem, this work presents a framework that uses parameter sets to define the most suitable update option for a particular business model. So data warehouse can implement a hybrid synchronism model where its portions synchronize at different time intervals depending on the characteristics of the transaction environment (information source) as well as the analytical application requirements.

Since data warehouse systems make use of storage techniques for efficient end user accessing and query facilities, these applications have implemented classic data synchronism operations that do not support an immediate data update. With the evolution of semantic data representation in the operational database environment, the accomplished analysis in data warehouse demands new synchronism ways. Hence, there is a growing interest in data warehouse system that can rapidly absorb the operational database updates, without compromising the operational query processes. Our research aims at the characterization of the synchronous and asynchronous algorithms limits for data updating in data warehouse systems. Our research proposes another way for update propagations of asynchronous transactions and details about these results can be seen in [Italiano and Ferreira 2003; 2006; Gonçalves et al. 2005].

7. AUTOMATION OF CLINICAL AND MOLECULAR DATA PROCESSING

Developing applications for e-Science presents many challenges. Some of them are related to the evolution of data requirements, other challenges arise with the advancement of scientific knowledge

which most often involves the change of requirements collecting and processing data. Furthermore, studies of software engineering show how costly is the process of changing requirements for software deployed. The DATA group have addressed this challenges developing software which can be easily adapted by the researcher, without knowledge about computing programming and with the guarantee of correctness and quality of results. Following, we present three examples of these software and its requirement challenges. The first one is the CEGH system that supports the changes in the human genomic tests in order to offer to patients results updated with the most recent genetic knowledge. The second is the DBCollHIV software developed to offer a collaborative environment to collect and analyze data about HIV. In 2010, the HIVdag was adopted by members of the PAHO-Pan American Health Organization, such as Uruguay, Argentina, Chile, Peru, Guatemala and some Caribbean countries. The third is a clinical database system available at http://clinmaldb.usp.br that was an initiative to integrate and automate the data collecting process on many sites of Amazon region.

All these softwares are based on DATA group results in the processing issues that have shown in above Sections and also in biological workflow requirements [Oikawa et al. 2004], database modeling techniques such as database modularization [Ferreira and Busichia 1999], naked objects framework to explore data requirements [Broinizi et al. 2008], and evolutionary database modeling [Domingues et al. 2009]. All these results in database modeling techniques has an evolutionary database modeling perspective that requires frequent changes in database application. This challenge is greater when the database must support multiple applications simultaneously. The current solution for evolution is the refactoring with a transition period. In this period, both the old and the new database schemas coexist and data is replicated through a synchronous process. This solution brings several difficulties, such as interference with the normal applications operation. To minimize these difficulties, we are using modularization [Ferreira and Busichia 1999], naked objects concepts [Broinizi et al. 2008], and asynchronous updates [Domingues et al. 2009] under integrated perspective to design, implement and maintain our database information systems.

7.1 Genoma Application - CEGH SYSTEM

The Human Genome Research Center (CEGH in Portuguese) at the University of São Paulo is the largest center in Latin America dedicated to the study of human Mendelian genetic disorders. Since its foundation about 40 years ago, more than 100,000 patients and their relatives have been referred to and examined by the different research groups. The main goal of the CEGH is to enhance comprehension of gene function with focus on neuromuscular, craniofacial, and brain development through the study of genetic disorders. The CEGH offers around 40 different genetic tests, which are being performed by several technicians under the supervision of 6 researchers. All samples to be analyzed have to be registered and sent to specialized technicians for analysis. It is crucial to have a flawless control-flow for each sample during every step.

Genetic disorders are very heterogeneous, which means that the phenotype can be caused by several genetic mechanisms involving more than one gene. Genetic tests are needed for precise clinic estimates of the recurrence genetic risks. To minimize testing costs, the strategy is to perform the genetic tests that account for the majority of cases for the studied disorder. Once a negative result is obtained, a second test is performed, which would account for the second most frequent cause of the disorder. At times, three or more tests need to be performed. Besides the heterogeneity of diseases, there are several mutations in gene that can cause the disease. For example, so far, there are more than 1500 mutations described for cystic fibrosis, which is the most common autosomal recessive disorder in Caucasians. As previously described, testing is started in regions most likely to have mutations; other segments of the gene are tested only if negative results were obtained previously.

This scenario requires software to handle several different views of procedures and tests. Moreover, users have to be guided to execution of the right task at the appropriated time, and the definition of new tests has to be a flexible routine. DATA group tools have been applied in this environment

creating CEGH Information System and in its essence, tests are defined as a set of actions, called procedures. The execution of these procedures will perform the desired genetic analysis. In addition, a procedure describes the techniques applied in each test's steps, and it contains all required information to perform the complete procedure; e.g., the list of reagents used.

CEGH Information System uses processes representation to manage the execution of genetic tests. This system has around one hundred tables and fifty interfaces to manage the data of patients and their families, doctors, lab's employers, diseases, clinical annotations, access control, tests, relations with tests and diseases, the order of tests, their execution and analyses. This System has been implemented with Ruby on Rails version 2.3 and uses a PostgreSQL database version 8.4 under a Linux environment. Details about the CEGH system can be seen online at: http://zen.genoma.ib.usp.br/. Although the system is implemented in Portuguese, we have provided an English translation together with a demo.

7.2 Clinical and Genoma: HIV System

Understanding the genetic diversity of HIV-1 and its biological consequences is important for designing effective control strategies. The improvement of sequencing technology has greatly increased the capacity of generating sequence data. With the widespread use of antiretroviral compounds against HIV, virus drug resistance has also become an important issue. Genotype testing for HIV drug resistance has been proved beneficial to treatment and now it is considered a standard of care procedure for individuals failing antiretroviral treatment. In Brazil, efforts were made to organize a network of laboratories (RENAGENO) to perform genotyping on patients failing therapy, and around 15,000 sequences are expected to be generated yearly. Other cohort studies are under ways that combine clinical data with viral genome sequences. Databases that manage sequences together with annotation are available for HIV and can be accessed through web interfaces. These databases are extremely useful for extracting reference sequence alignments. Nevertheless, the access to sequence information and clinical data from the patient is still very limited. Furthermore, an important feature for the investigator is to be able to manipulate his/her raw data before publication. The development of a database system based on WED-flow approach that is able to handle and search sequences produced locally and to integrate them with epidemiological and clinical data and tools of bioinformatics, would provide a major advantage for research groups in developing countries that do not have the necessary resources to develop their own systems.

The application named Database System for Collaborative HIV analysis (DBCollHIV) [Araújo et al. 2006] has been developed to manage sequence, clinical, epidemiological and treatment data generated by ongoing HIV studies in Brazil, such as [Barreto et al. 2006] (http://clinmaldb.usp.br/dbcollhiv). DBCollHIV was projected for storing information related to clinical evolution of HIV treated patients. It also provides tools for analyzing sequences as HIV subtype analysis [Araújo et al. 2009], drug resistance tool [Araújo et al. 2008] which evaluates the relationship between mutations and drug resistance, using the rules established by the Ministry of Health - Committee of Experts RENAGENO [12]. It also offers tools for sequence analyses such as hiv subtype analysis [Araújo et al. 2009], drug resistance tool for interpretation of relation between mutations and drug resistance, using rules created by the Brazilian Ministry of Health RENAGENO Expert's Committee [13] and used to develop an algorithm to automate the process [Araújo et al. 2008]. These tools enable automatic analysis of two important regions of the HIV genome, reverse transcriptase and protease, and its results are used by physicians to support the decision to change the patients treatment. Furthermore, the researchers can change the rules for evaluation of mutations, automatically generate the new algorithm for identification of resistance to antiretroviral drugs and to compare the impact of new rules [14].

[12] http://www.aids.gov.br/ renageno.htm.
[13] http://www.aids.gov.br/ renageno.htm
[14] http://clinmaldb.usp.br:8083/hiv/resistencia/resistencia.html.

7.3 Epidemiology and Clinical Information: ClinMalDB Database System

Despite many research efforts, malaria remains one of the most serious parasitic diseases of the world, responsible for millions of deaths, mainly children below five years old. Some previous works towards that malaria cases in Brazil have special characteristics, related to other endemic regions of the world. The Brazilian number of cases is also considerable, involves a particular social-economical context and two important species *Plasmodium falciparum* and *Plasmodium vivax*, offering an ideal environment for comparative studies. Based on this favorable research opportunity, DATA group worked in cooperation with the research group headed by Prof. Hernando A. del Portillo on studying diverse clinical and epidemiological aspects of human malaria in Brazilian Amazon region.

The main computational result of this work was the definition of a clinical database system (ClinMaldDB). ClinMalBD system[15] [Oikawa et al. 2010] was an initiative to integrate and automate data collecting process on many sites of Amazon region. This system considers the creation of an integrate environment and a single interface for all local research and medical groups interested on unifying the collecting process of clinical data, epidemiological data and physical samples. After diagnosis and data collecting, replacing printed paper forms, samples were sent to specific laboratories, for further work in Molecular Biology studies. ClinMalDB provides a platform for storing, manipulating and pre-analysing information, since it maintains relationships among the main identification of patients, doctors, samples, freezers (local sample repositories), clinical and epidemiological data.

Under structure point of view, ClinMalDB was developed using a modularization approach and four levels with different purposes. The first level is responsible for primary data storage; the second one is composed of integration routines for the databases inside the first level; level three has cube views; and level four houses all components responsible for user data retrieval including components for data visualization, data entry, and data export.

Under database logical point of view, the data organization was built using semantic modules, in order to facilitate future expansions and integration with correlate software. The main modules of last database version were:

—Patients: contains information registered in the Health Centers by the physicians assisting malaria patients including ethnicity information.
—Exams: contains information registered in the Health Centers by the physicians assisting malaria patients including clinical and epidemiological data.
—Samples: contains information related to the manipulation and physical storage of the blood samples collected from registered patients.
—Sequences: contains genomic and cDNA sequences from parasite coding genes submitted to automated quality assessment. This information is usually provided by static databases.
—Users: contains the registered set of users of the system along with a description of which information they can access and/or manipulate.

This project is considered finished and was responsible for publication of diverse papers on conferences and journals, covering many research areas, such as Molecular Biology, Biology, Bioinformatics and Computers Science such as [Albrecht et al. 2006; Merino et al. 2006]

8. NEXT STEPS FOR DATA GROUP

Our current research plan includes the following next steps in: transaction processing; performance analysis for business processes and scientific workflows; dimensionality reduction and classification of time series in data warehouses; and automation of clinical and molecular data processing.

[15]available at http://clinmaldb.usp.br

8.1 Transaction Processing

8.1.1 *Exception Handling Problem.* One of the major challenges in the delivery of high quality services is the handling of the many exceptional cases in complex applications. Due to the lack of use, the code for exception handling is often insufficiently tested or out of date. At the same time, the quantity of exception handling code is proportional to the complexity of such cases, which remains high. As result, such exception handling is in practice delegated to human operators in call centers, causing significant response delays and high costs.

8.1.2 *Mission-Critical Services.* Some important categories of applications are considered mission-critical, i.e., their failure would cause significant losses and/or damages to customers and providers. Examples of large scale mission-critical applications include financial, military, and e-commerce. These important applications typically have real-time quality of service requirements (e.g., limited response time and high availability) that are difficult to satisfy in a call center environment staffed by cold-start remote operators.

8.1.3 *Limitations of Process-Oriented Approaches.* One of the classic approaches to service composition (and workflow orchestration) is based on directly translating manual business processes into programmed automated processes. This automation through translation is effective for the "normal path", where the execution follows the expected outcome. However, as an application becomes more popular and successful, it also becomes more sophisticated and complex. This growth is often due to the many additional cases of exceptions beyond the "normal path" that need to be considered. The usual practice of writing exception handling code in procedural programming languages leads to exponential state and code explosion, since each one of these exceptions often requires similar state and code size as the normal path. In summary, it is difficult to develop the exception handling code; it is difficult to test and validate it; it is difficult to maintain it and evolve it when the business processes change.

8.2 Performance Analysis for Business Processes and Scientific Workflows

The main limitation of the use of analytical models to predict the performance of business processes and scientific workflows is the problem of the state space explosion of Markovian formalisms. Even small parallel/distributed systems can generate large state spaces due to the complexity of their behavior and resource requirements. It is well-known that the business process and the scientific workflow models can be constituted of a large number of tasks, combined by very sophisticated structures of control-flow. This complex behavior may imply in analytical models with huge state spaces, whose numerical solution can require prohibitive computational efforts (both in terms of processing time and memory consumption).

For this reason, the study and the development of techniques to improve the "analyzability" of the Markovian models is a mandatory step if we want to apply performance analysis in large-scale models. Exploring the particularities of the domain, we will be able to propose decomposition methods that enable us to use a divide-and-conquer approach in the analysis of business process/workflow models with large state spaces.

8.3 Dimensionality Reduction and Classification of Time Series in Data Warehouses

One of the characteristics that differentiate a data warehouse from other systems is that in the former data is associated to time information and to established time periods. In a multidimensional model, it is very rare to find a dimensional model that does not contain time as a fundamental dimension. In this way, all the information contained in a data warehouse is related to time, as every line stored

in the fact table contains a key that relates it to the dimension of time. It is worth emphasizing that a multidimensional model can contain several time dimensions. Hence, every metric that is related to a time dimension will result in a time series that can be analyzed. A search for the measurement hemoglobin, when analyzed together with the dimension date of exam, will result in time series of hemoglobin by exam date for each patient. For each new dimension included in the search, and for each element present in the hierarchy of the selected dimension, many time series will be created. This implies an exponential growth in the number of time series present in a data warehouse. Therefore, it is important to develop computer techniques that automatically identify and reduce the dimensions that are not significant to the time series stored in the data warehouse.

The search for similarity in time series is another challenge in data warehouse systems. Finding similar time series in a multidimensional database is a task that requires great computational effort. Hence, researching and adapting algorithms that scan the entire multidimensional database in search of similarities or abnormalities is still an area under development. Techniques such as the cleaning and removal of "white noise", the supervised and unsupervised recognition of patterns (including clustering), the transformation and analysis of wavelets and other multi-scale techniques may be adapted to accelerate the identification and characterization of time series.

8.4 Automation of Clinical and Molecular Data Processing

The automation of clinical and molecular data processing enables fast adaptation of systems to the new requirements introduced by the advance of scientific knowledge. However, these scientific insights can guide to changes into stored results. In some cases, these changes can improve or even invalidate the previous results. This scenario shows the automation of clinical and molecular data processing must not only ensure that new processes can be adjusted accurately and automatically by the user, but also offers resources for managing the evolution of analysis performed and results obtained. Accordingly, the DATA group has developed research in the area of Data Provenance to provide resources for monitoring and evaluation of the transformations performed in the data.

9. CONCLUSIONS

In the last 11 years the DATA group has graduated 23 master degree students and 6 PhD students and support 12 students in undergraduate research. In terms of publications and international impact, we have generated 45 papers in conference events and 11 journal papers in issues described in above Sections. In terms of grant research and multidisciplinary project we have received significant financial resources to support Data group laboratories and infra-structure as well as student grants. Our automation of clinical and molecular data processing researches has a local and regional impacts covering: 1) in the case of blood donation system our plan is to move to the integration of more blood centers in order to gradually cover all blood collections in Brazil. Currently, our REDS-II data warehouse stores above 1.5 million screenings from about 860,000 candidates, with more than 1.1 million donations and 340,000 deferrals, and from the blood samples, more than 9.5 millions tests have been performed; 2) in the HIV System all reports and records about HIV patients in Brazil, in which the HIVdag analyzes 15,000 exams per year and furthermore, the DBCollHIV has been used by Brazilian Health Ministry to manage projects in HIV molecular analysis; 3) in the CEGH system stores 100,000 patients and 2,500 molecular exams and visits per year.

With the results above reported, our DATA group shows its capability to support research initiatives and to offer technologic transaction processing tools able to be used in real database information systems. Our research conducting strategy in transaction processing for real applications (but not restricted to them) allows us to know concrete problems and challenges, enabling us to make contributions in modern science, that is getting more interdisciplinary and data-intensive. Our ongoing steps include data mining and information retrieval to support e-Science applications described in Section 7.

REFERENCES

Albrecht, L., Merino, M. F., Hoffmann, E. H., Ferreira, M. U., de Mattos Ferreira, R. G., Osakabe, A. L., Martha, R. C. D., Ramhrter, M., Durham, A. M., Ferreira, J. E., and del Portillo, H. A. Multi-character population study of the vir subtelomeric multigene superfamily of plasmodium vivax, a major human malaria parasite. *Molecular and Biochemical Parasitology* 149 (1): 10–16, 2006.

Almeida-Neto, C., Lui, J., Wright, D. J., Mendrone-Junior, A., Takecian, P. L., Sun, Y., Ferreira, J. E., de Alencar Fischer Chamone, D., Bush, M. P., and Sabino, E. C. Demographic profile of blood donors at three major brazilian blood centers: results from the international reds-ii study, 2007 to 2008. *Transfusion (Arlington, Va)* 51 (1): 191–197, 2009.

Araújo, L. V., Sabino, E. C., and Ferreira, J. E. Hiv drug resistance analysis tool based on process algebra. In *Proceedings of the 2008 ACM symposium on Applied computing*. SAC '08. ACM, New York, NY, USA, pp. 1358–1363, 2008.

Araújo, L. V., Sanabani, S. S., Sabino, E. C., and Ferreira, J. E. Hivsetsubtype: software for subtype classification of hiv-1 sequences. In *Proceedings of the 2009 ACM symposium on Applied Computing*. SAC '09. ACM, New York, NY, USA, pp. 811–815, 2009.

Araújo, L. V., Soares, M. A., Oliveira, S. M., Chequer, P., Tanuri, A., Sabino, E. C., and Ferreira, J. E. Dbcollhiv: A database system for collaborative hiv analysis in brazil. *Genetics and Molecular Research* 3 (1): 203–215, 2006.

Barreto, C. C., Nishyia, A., Araújo, L. V., Ferreira, J. E., Busch, M. P., and Sabino, E. C. Trends in antiretroviral drug resistance and clade distributions among hiv-1–infected blood donors in sao paulo, brazil. *Journal of Acquired Immune Deficiency Syndromes* 41 (1): 388–341, 2006.

Braghetto, K. R. *Padrões de Fluxos de Processos em Banco de Dados Relacionais*. M.S. thesis, Instituto de Matemática e Estatística - Universidade de São Paulo, 2006.

Braghetto, K. R., Ferreira, J. E., and Pu, C. Using control-flow patterns for specifying business processes in cooperative environments. In *Proceedings of the 2007 ACM Symposium on Applied Computing*. SAC '07. ACM, Seoul, Korea, pp. 1234–1241, 2007.

Braghetto, K. R., Ferreira, J. E., and Pu, C. Business processes management using process algebra and relational database model. In *Proceedings of the International Conference on e-Business*. ICE-B 2008. INSTICC Press, Porto, Portugal, pp. 323–333, 2008a.

Braghetto, K. R., Ferreira, J. E., and Pu, C. Using process algebra to control the execution of business processes. In *Proceedings of the 2008 ACM Symposium on Applied Computing*. SAC '08. ACM, New York, NY, USA, pp. 128–129, 2008b.

Braghetto, K. R., Ferreira, J. E., and Pu, C. NPTool: Towards scalability and reliability of business process management. In *e-Business and Telecommunications*, J. Filipe and M. S. Obaidat (Eds.). Communications in Computer and Information Science, vol. 48. Springer Berlin Heidelberg, pp. 99–112, 2009.

Braghetto, K. R., Ferreira, J. E., and Vincent, J.-M. Performance analysis modeling applied to business processes. In *Proceedings of the 2010 Spring Simulation Multiconference*. SpringSim '10. ACM, New York, NY, USA, pp. 122:1–122:8, 2010.

Braghetto, K. R., Ferreira, J. E., and Vincent, J.-M. From Business Process Model and Notation to Stochastic Automata Network. Technical Report (Reference Number: RT-MAC-2011-03), University of São Paulo. Mar., 2011. [Available at: `http://www.ime.usp.br/~kellyrb/files/fromBPMNtoSAN.pdf`].

Broinizi, M. E. B., Ferreira, J. E., and Goldman, A. Using annotations in the naked objects framework to explore data requirements. In *SAC*. pp. 630–637, 2008.

Cantão, M. E., Araújo, L. V., Lemos, E. G. M., and Ferreira, J. E. Algebraic approach to optimal clone selection applied in metagenomic projects. In *Proceedings of the International Symposium on Biocomputing*. ISB '10. ACM, New York, NY, USA, pp. 36:1–36:6, 2010.

Carneiro-Proietti, A. B., Sabino, E. C., Sampaio, D., Proietti, F. A., GonÐalves, T. T., Oliveira, C. D. L., Ferreira, J. E., Liu, J., Schreiber, G. B., Murphy, E. L., and Bush, M. P. Demographic profile of blood donors at three major brazilian blood centers: results from the international reds-ii study, 2007 to 2008. *Transfusion (Arlington, Va)* 149 (1): 20–32, 2009.

Domingues, H. H., Kon, F., and Ferreira, J. E. Replicação assíncrona em modelagem evolutiva de banco de dados. In *SBBD*. pp. 121–135, 2009.

Ferreira, J. E. and Busichia, G. Database modularization design for the construction of flexible information systems. In *IDEAS*. pp. 415–422, 1999.

Ferreira, J. E., Takai, O. K., Braghetto, K. R., and Pu, C. Large scale order processing through Navigation Plan concept. In *IEEE International Conference on Services Computing, 2006 (SCC'06)*. IEEE Computer Society, Chicago, USA, pp. 297–300, 2006.

Ferreira, J. E., Takai, O. K., Malkowski, S., and Pu, C. Reducing exception handling complexity in business process modeling and implementation: The wedflow approach. In *Proceedings of 18th Conference on Cooperative Information Systems (CoopIS)*. LNCS. Springer, Crete, Greece, pp. 150–167, 2010.

Ferreira, J. E., Takai, O. K., and Pu, C. Integration of business processes with autonomous information systems: A case study in government services. In *CEC*. pp. 471–474, 2005a.

Ferreira, J. E., Takai, O. K., and Pu, C. Integration of collaborative information system in internet applications using riverfish architecture. In *CollaborateCom*, 2005b.

Ferreira, J. E. and TrainaJr, C. Adding conceptual constructs to support distribution in an object oriented model. In *IX SimpŮsio Brasileiro de Banco de Dados*. SBBD1994. SBC, Sao Carlos, Sao Paulo - BR, pp. 143–157, 1994.

Ferreira, J. E. and TrainaJr, C. Controle de compartilhamento e acesso em gbdoo baseado em composicao de objetos. In *XI Simposio Brasileiro de Banco de Dados*. SBBD1996. SBC, Sao Carlos, Sao Paulo - BR, pp. 143–157, 1996.

Ferreira, J. E., Wu, Q., Malkowski, S., and Pu, C. Towards flexible event-handling in workflows through data states. In *Proceedings of 4th International Conference on Scientific Workflows*. Miami, FL, USA, pp. 344–351, 2010.

Garcia-Molina, H. and Salem, K. Sagas. In *Proceeding of ACM SIGMOD Conference on Management of Data*. San Francisco, CA, USA, pp. 249–259, 1987.

Gonçalves, B. M. M. T., Italiano, I. C., and Ferreira, J. E. Data updating between the operational and analytical databases through dw-log algorithm. In *IDEAS*. pp. 77–82, 2005.

Italiano, I. C. and Ferreira, J. E. A hybrid model for data synchronism in data warehouse projects. In *IDEAS*. pp. 12–21, 2003.

Italiano, I. C. and Ferreira, J. E. Synchronization options for data warehouse designs. *IEEE Computer* 39 (3): 53–57, 2006.

Merino, E. F., Fernandez-Becerra, C., Durham, A. M., Ferreira, J. E., Tumiasci, V. F., dArc Neves, J., Dadario, A., Silva-Nunes, M., Ferreira, M. U., Wickramarachchi, T. A., and del Portillo, H. A. Extense variant gene family repertoire overlap in western amazon plasmodium falciparum isolates. *Molecular and Biochemical Parasitology* 150 (2): 157–165, 2006.

Oikawa, M. K., Broinizi, M. E. B., Dermagos, A., Armelin, H. A., and Ferreira, J. E. Genflow: Generic flow for integration, management and analysis of molecular biology data. *Genetics and Molecular Biology* 24 (4): 691–695, 2004.

Oikawa, M. K., Ferreira, J. E., Malkowski, S., and Pu, C. Towards algorithmic generation of business processes: From business step dependencies to process algebra expressions. In *International Conference on Business Processes Management (BPM'09)*. LNCS. Springer, Ulm, Germany, pp. 80–96, 2009.

Oikawa, M. K., Merino, F. E., Wunderlich, G., Villalobos, J. M., Silva, P. P., Barrera, J., Durham, A. M., del Portillo, H. A., and Ferreira, J. E. Clinmaldb: A clinical field-research oriented relational database to study human malaria. *IADIS International Journal on WWW/Internet* 8 (4): 181–191, 2010.

Rodrigues, M. C., Malkowski, S., and Ferreira, J. E. Implementing rigorous web services with process algebra: navigation plan for web services. In *Proceedings of the 2009 ACM Symposium on Applied Computing*. SAC '09. ACM, New York, NY, USA, pp. 625–631, 2009.

Takecian, P. L. *ACP e LOTOS: um estudo comparativo baseado em conceitos de BPEL e padrões de controle de fluxo*. M.S. thesis, Instituto de Matemática e Estatística - Universidade de São Paulo, 2008.

Takecian, P. L., Ferreira, J. E., Malkowski, S., and Pu, C. Using LOTOS for rigorous specifications of workflow patterns. In *Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), 2010 6th International Conference on*. IEEExplorer, Chicago, USA, pp. 1–7, 2010.

TrainaJr, C., Ferreira, J. E., and Biajiz, M. Use of a semantically grained database system for distribution and control within design environments. In *Euro-Par*. pp. 1130–1134, 1997.

Wu, Q., Pu, C., and Ferreira, J. E. A partial persistent data structure to support consistency in real-time collaborative editing. In *ICDE*. pp. 776–779, 2010.

Zuliane, D., Oikawa, M. K., Malkowski, S., Alcazar, J. J. P., and Ferreira, J. E. The riverfish approach to business process modeling: Linking business steps to control-flow patterns. In *CollaborateCom*. pp. 179–193, 2008.