Extending scientific workflows for managing hypotheses and models

Bernardo Gonçalves, Fabio Porto, Ana Maria de C. Moura

Extreme Data Lab (DEXL Lab)
Laboratório Nacional de Computação Científica (LNCC)
Petrópolis – RJ – Brasil

{bgonc,fporto,anamoura}@lncc.br

Abstract. *Support for in-silico scientific exploration has been focusing on the research experimental phase. Nevertheless, there is valuable information associated to the research endeavor that has been left out of reach of the eScience environment and is mostly kept by scientists tacitly or in unstructured forms in their desktop environment. It concerns a description of the observed phenomenon, the conceptual formulation of scientific hypotheses as tentative explanations for it, as well as the models developed to encode these hypotheses. In this paper, we propose to extend scientific workflows in order to include such entities as data elements in the eScience environment, enabling then for their semantic management. These elements are integrated in a three-layered architecture which is illustrated in a case study in the modeling and simulation of the human cardiovascular system.*

1. Introduction

New instruments and technologies used during *in-silico* experiments are producing an enormous amount of scientific data. In projects in which our group is involved, we are managing hundreds of terabytes of astronomy data reduced from telescope images of the Dark Energy Survey Project (DES)[1], and tens of terabytes from simulations of the human cardiovascular system produced by the Hemolab simulation environment [Blanco et al. 2000]. In scientific endeavors of this magnitude, the scenario of a scientist running *ad-hoc* programs in his/her desktop as a scientific platform is getting rare. In large-scale science, a High-Performance Computing (HPC) infrastructure is used to run scientific workflows that prepare and analyze the data, in most cases, using some data partitioning and parallel execution strategies *à la* MapReduce [Dean and Ghemawat 2004]. Scientific workflows can also be equipped with provenance management to add reproducibility and traceability to the experiment execution, as it is the case of systems like Vistrails [Bavoil et al. 2005], Chiron [Ogasawara et al. 2011], and QEF [Porto et. al 2007].

An eScience environment with such capabilities supports the scientific *experiment* life-cycle [Mattoso et al. 2010]. In this paper, we contribute to the problem of supporting *in-silico* scientific research by extending its frontiers to the complete scientific research life-cycle, from the observed phenomenon to hypothesis formulation to its experimental evaluation and validation against data. In this broader context, the experiment composition and evaluation is only one (yet essential) stage of the process. Currenlty, a considerable amount of provenance information associated to the research endeavor is left out of the reach of the eScience environment and is kept by scientists tacitly or in unstructured forms in their desktop environment. In large collaborative projects, such as DES and LSST (Large Synoptic Survey Telescope), understanding the phenomenon as seen by the modeler(s) is paramount for people in collaboration to interpret their colleagues' results. Moreover, the process of understanding the observed phenomenon itself evolves. It turns out that this wealth of provenance information requires a data-oriented approach that should accommodate different elements of the scientific research life-cycle into a unified conceptual framework.

In this context, we propose a three-layered conceptual architecture that covers: (i) the observed phenomenon, (ii) the scientific hypotheses, (iii) the models that encode the hypotheses, (iv) the scientific workflows that evaluate the hypotheses in experiments, and (v) the data that both feeds and is produced by the experiment. We have separated these elements in three layers

1 DES-Brazil,The Dark Energy Survey Project, Brazil, http://des-brazil.linea.gov.br/.

in order to reflect the practice of the scientist. These layers are the conceptual, logical and physical, as usual in the theory and practice of databases. They are aimed at addressing the representation and management of the data elements just mentioned. The conceptual layer addresses the phenomenon description and the hypothesis conceptual formulation. The logical layer in turn copes with the models that express the hypotheses in some formal language, and/or the computational procedure that simulates it.[2] Finally, the physical layer deals with the data issues and the workflow execution of the experiment. Thus, both the conceptual and logical layers compose the extension proposed in this paper. This extension has been conceived in such a way that the complete process involved in the scientific endeavor can be accessed and the evolution of the phenomenon understanding can be traced back. We illustrate the proposed architecture in a research aiming at predicting the behaviors of the human cardiovascular system through computational simulations. For brevity, we focus that illustration on the original (extended) elements we have been referring to, viz., the conceptual hypotheses and their associated models.

The remainder of this paper is organized as follows. In Section 2, we discuss related work. Section 3 presents the proposed three-layered architecture. Next, in Section 4, the case study in the modeling and simulation of the human cardiovascular system illustrates the original elements of the architecture and their relations. Finally, Section 5 presents our conclusions.

2. Related Work

Back in the 80's, the notion of hypothetical databases emerged in the context of database management systems [Bonner 1990]. They, however, were not data models for scientific hypotheses, but hypothetical states for arbitrary databases. These states were produced by delete and insert operations, and queries that could be satisfied on such state. Rather, we are addressing the conceptual modeling of scientific hypotheses for data and knowledge engineering. This is in fact a barely explored *conceptual* problem. Under that perspective, we refer in the following to a research initiative that has appeared in the last decade in Bioinformatics.

The conceptual framework of HyBrow (Hypothesis Browser) [Racunas et al. 2004] aims at providing scientists with a unified eScience infrastructure for both hypothesis formulation and evaluation against observational data in Molecular Biology. HyBrow employs an OWL ontology and application-hardcoded rules for inference from facts stored in an integrated knowledge base. HyQue [Callahan et al. 2011] is in turn an adaptation of HyBrow for the linked data technologies RDF/SPARQL, which adds to it semantic interoperability capabilities and leverages to some extent its conceptual expressivity.

HyBrow/HyQue's hypotheses are formalism-specific assertions forming a set H in a Knowledge Base (KB). The KB has also rules that model accepted assertions over the same universe and experimental data. The KB then can contradict or validate some of the hypothesis statements, leaving others as candidates for new discovery. As more experimental data is obtained and new rules are inserted, discoveries either accumulate evidence or are contradicted. In the latter case, the correlated rules must be identified and eliminated from the theory H. The hypotheses in H correlate biological processes (seen as events) are represented in FOL with free quantifiers (see an example below).

HyBrow/HyQue's Hypothesis (from [Callahan et al. 2011]):

e1 (Gal4p induces expression of GAL1) OR
  e2 (Gal3p induces expression of GAL2
  e3 AND Gal4p induces expression of GAL7) OR
  e4 (Gal4p induces expression of GAL7
  e5 AND Gal80p inhibits production of Gal4p
  when GAL3 is over-expressed
  e6 AND Gal80p induces expression of GAL7)

We consider the HyBrow/HyQue framework as lying at the logical and physical layers (see Section 3). They assume a formalism-specific encoding to the hypothesis statements, in a conceptual framework meant for both hypothesis formulation and evaluation. Instead, we are looking also at the scientists' hypotheses conceptually and dealing with the challenge of managing them semantically at the conceptual level. At the logical level, the hypotheses can

2 For instance, the mathematics of the differential and integral calculus used to express a continuous view of natural phenomena, and a scientific programming language used to simulate the phenomenon; or a computer theoretic formalism like Petri Nets to describe (say) a message-passing mechanism of protein synthesis.

then be encoded in proper machine-understandable formalism (like in HyBrow) to be evaluated *in silico* at the physical level in a scientific workflow.

3. The Three-Layered Architecture

As previously mentioned, our proposed architecture comprises three layers, viz., the traditional separation in conceptual, logical and physical layers, see Fig. 1. When we speak of the issues to be addressed in this architecture, we are always referring to data representation and management issues. A preliminary version of data models for the entities (i) Observed Phenomenon, (ii) Scientific Hypothesis, and (iii) Model (mathematical or computational) have been developed and can be found elsewhere [Porto et al. 2008; Porto and Spacappietra 2011]. Those data models are being elaborated in such a way that it will be possible to track the conceptual and logical scientific modeling life-cycles. In this paper, we are abstracting from particular data representations to focusing on how such elements relate to each other in the proposed conceptual architecture for eScience. Thus, scientific hypotheses and models expressing them are abstracted in Section 4 as data elements represented, respectively, by variables $h_1, h_2, \ldots, h_n$, and $m_1, m_2, \ldots, m_n$, with $n \in \mathbb{N}$. Data management in turn is characterized at the novel conceptual and logical layers as semantic management, as illustrated in Section 4.

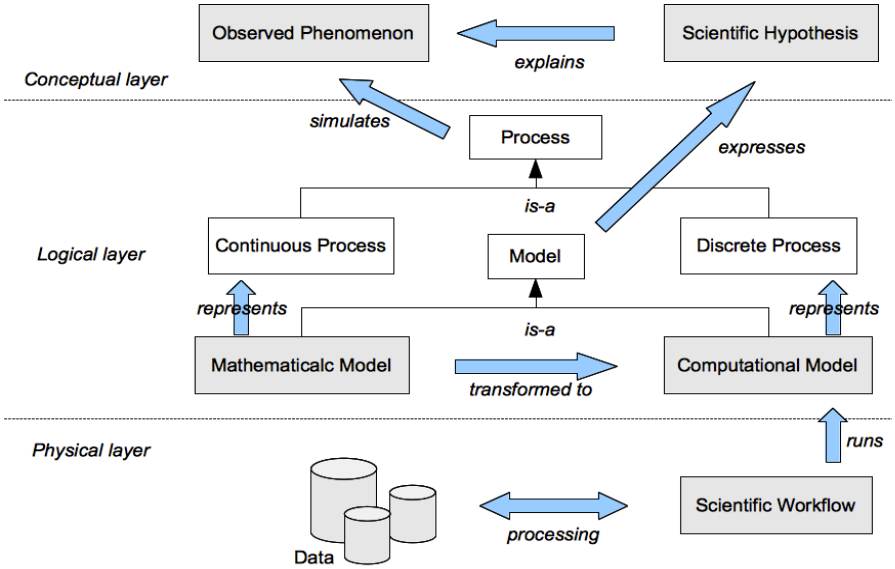

Fig. 1: Three-layered conceptual architecture for eScience.

The conceptual layer comprises a description of the observed phenomenon and its associated explanations, the scientific hypotheses. For the purpose of this paper, the phenomenon description can be understood as being composed of a unique identifier, a set of observables (e.g., mathematical variables *x*, *y* and *z*), and an enunciation like "*population dynamics*", "*blood flow in a cardiovascular vessel*", or "*genesis of tsunamis*".

A scientific hypothesis is a statement of explanation that must be refutable [Popper 2002]. A model *expresses* a scientific hypothesis in some formal language (e.g., mathematics). We assume that every (scientific) model expresses exactly one hypothesis, and formalize this relationship by a function $\upsilon : M \to H$, where *M* is a set of models, and *H* is a set of hypotheses. Models embodying hypotheses play then a fundamental role in experimental science by bringing rigor to problem statement and results validation. In the *in-silico* research, experiments support or refute hypotheses through simulations whose results establish a distance between the hypothesis and the associated observations. This allows one to assign quantitative values to hypotheses (semantic entities), bridging then the gap betwen the quantitative and qualitative views. Unlike the HyBrow approach, we consider that notion of pragmatics, i.e., that hypotheses encoded in models approximate the observed phenomena according to some metrics. Therefore, we define for each hypothesis the refutability property, which is a continuous function $\rho : H \to [0,1]$. Notice that a composite function $\rho \circ \upsilon : M \to [0,1]$ then retrieves the refutability of a model. This conceptual modeling of scientific hypotheses introduces a valuable

contribution by bridging the gap between qualitative description of the phenomenon domain and the corresponding quantitative valuation.

At the logical layer the observed phenomenon is represented as a spatio-temporal Process, as proposed by Sowa [2000]. On the one hand, the continuous view of phenomena in space-time is represented by a set of integral or differential equations that model some Continuous Process. On the other hand, a Discrete Process represents the state-transition view of the phenomenon simulation, as reproduced in a mathematical/logical discrete-event model or in a computational model. A Mathematical Model is eventually *transformed to* a Computational Model. The idea of such an integrated view under the Process rubric is to keep those different elements of the *in-silico* research still referring to the studied phenomenon as an anchor to the investigation. At the logical layer, models *simulate* the observed phenomenon.

Finally, the physical layer deals with the representation and management of the experiment, as approached by scientific workflows and their processing data - input data, as well as the data produced during the experiment runs and all the provenance data gathered during experiment evaluation. The physical layer is not considered in detail in Fig. 1, neither is it discussed thoroughly in this paper. For an in-depth account of the representation and management issues at this layer, one can refer to [Mattoso et al. 2010].

By establishing those three layers and their interfaces, scientists have then an explicit reference of the hypotheses formulated, the models expressing them and their associated experiments, all together in the complete scientific life-cycle (see Fig. 2). From the traditional experiment point of view, new kinds of provenance information arises concerning models and hypotheses. Interesting queries can be formulated, for example, to retrieve the hypothesis a given model expresses, or what hypotheses are near enough as explanations for the studied phenomenon. All this can be quite useful in a collaborative *in-silico* research. With this purpose, the implementation of the scientific hypothesis and model data elements in the semantic web/linked data technologies OWL/RDF shall enlarge the potential of applicability of the proposed three-layered architecture in the context of the Linked Data effort, in particular, in the emerging community of Linked Science.[3] We proceed now to the illustration of the conceptual and logical issues through an example of *in-silico* scientific research which is a representative example.

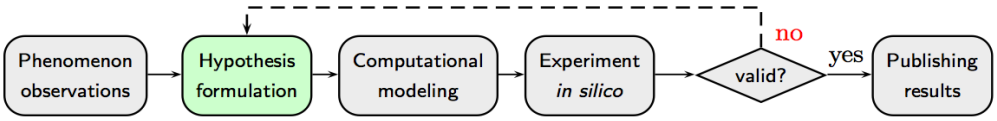

Fig. 2: Flow chart of a complete life-cycle of an *in-silico* research.

4. Case Study: Modeling the Human Cardiovascular System

In order to illustrate the data elements appearing in the two novel layers, viz., the conceptual and the logical layers, we refer to the modeling and simulation of the human cardiovascular system developed at LNCC[4], in support of the medical diagnosis of cardiovascular diseases. This example of scientific modeling activity starts with a simplistic representation of the human cardiovascular system, in which the parts of the system (a network of blood vessels) are modeled as lumped (non-spatial) physiological components [Liang et al. 2009; Blanco et al. 2009]. These components, the blood vessels, are seen by analogy as resistive-capacitive electrical circuits, hence the same physical laws (e.g., Ohm's law) hold for them. This is the so-called 0-D model, which comprises, mathematically, ordinary differential equations. The cardiovascular system is then modeled as a lumped (closed-loop) dynamic system. We have here a hypothesis about how a generic component of the cardiovascular system behaves, viz., that ($\boldsymbol{h}_1$) *"A blood vessel behaves as a lumped RC-circuit"*, which is expressed by mathematical model ($\boldsymbol{m}_1$), as shown in Table 1. Then a computational model (say, a scientific program coded in MATLAB) as a transformation of ($\boldsymbol{m}_1$) can simulate the observed phenomenon (i.e., the observed blood flow over the vessel). The computational model ($\boldsymbol{m}_1$*) can be linked to a workflow data element in order to be run, hence evaluating the hypothesis ($\boldsymbol{h}_1$) it expresses.

3 http://linkedscience.org/.

4 LNCC - National Laboratory for Scientific Computing; cf. the project INCT-MACC at http://macc.lncc.br.

Table 1: Instances of scientific hypotheses and instances of mathematical models that express them. The hypotheses lie at the conceptual layer, while the models lie at the logical layer. This table does not account for a precise data representation of such data elements.

Hypothesis Key	*Hypothesis Natural Language Statement*	*Model Key*	*Mathematical Model*
h_1	"*A blood vessel behaves as a lumped RC-circuit*".	m_1	$C_j \frac{dP_j}{dt} = Q_j - Q_{j+1}$ $R_j Q_j = P_{j-1} - P_j$
h_2	"*A blood vessel behaves as a lumped RLC-circuit*".	m_2	$C_j \frac{dP_j}{dt} = Q_j - Q_{j+1}$ $L_j \frac{dQ_j}{dt} + R_j Q_j = P_{j-1} - P_j$
h_3	*"A blood vessel behaves as a lumped RC-circuit and an external pressure is exerted on it."*	m_3	$C_j \frac{d(P_j - P_{ex})}{dt} = Q_j - Q_{j+1}$ $R_j Q_j = P_{j-1} - P_j$
h_4	"*The blood flows radially over the vessel [now, a pipe] as a viscous fluid*."	m_4	$v(r) = \frac{(P_{j-1} - P_j)}{4L\mu}(R^2 - r^2)$ $Q_j = \frac{\pi R^4}{8\mu L}\frac{(P_{j-1} - P_j)}{L}$
h_5	"*The blood behaves as a viscous fluid*."	m_5	$\rho\left(\frac{\partial \vec{v}}{\partial t} + \vec{v}\cdot\nabla\vec{v}\right) = -\nabla p + \nabla\cdot\vec{T} + \phi$

At this point, it is worthwhile recalling George Box's quoting, the motto of mathematical modeling, that "*all models are wrong, some are useful*." In fact, the 0-D model is a simplistic representation of the cardiovascular system, yet one which is the basis for more sophisticated models, and also one that is able to achieve a number of relevant predictions; e.g., to predict how the patient's cardiac output and systemic pressure will change over his/her aging, or even the calibration itself of anatomical and physiological parameters by taking canonical values of an average individual. Now, suppose that after a research life-cycle comprising the three central stages illustrated in Fig. 2, the distance between data produced by the simulation and data that has been observed is significant. Additionally, suppose it turns out that a model fine tuning (say, parameter recalibration) brings no sensible effect. The scientist might consider in this validation stage that hypothesis h_1 does not seem to hold. In our example, feasible hypothesis reformulations could be (h_2) "*A blood vessel behaves as a lumped RLC-circuit*" and (h_3) *"A blood vessel behaves as a lumped RC-circuit and an external pressure is exerted on it."* (see them associated to their mathematical models in Table 1). While the former considers inertial effects inhibiting the vessel wall to get deformed at its resting state, the latter considers the effect of an external pressure (e.g., due to breathing) being exerted on the vessel wall that constrains its deformation. The scientist then can recall his/her observations of the phenomenon in order to shed light on his/her hypothesis reformulation.

Our point w.r.t. integrating both conceptual hypotheses and models as data elements into the *in-silico* environment is then illustrated clearly now. Consider that data representations of hypotheses h_1, h_2 and h_3 are linked to their mathematical/computational model data representations m_1, m_2 and m_3, which are in turn linked to the results analysis module in the physical layer. Then we can provide the scientist with semantic management features to track his/her research likely in a way more efficient than before, outside the eScience environment. He/she will be able to query the refutability of his/her hypotheses (like we have mentioned in the previous section). Besides, modifications may take place at the logical layer, which are not necessarily derived from hypothesis reformulation. A typical activity of this sort, as mentioned above, is the so-called model tuning, where the scientist calibrates the computational model parameters without recurring to hypothesis reformulation.

Finally, notice that the simplistic modeling step that has been exposed in our example is meant to keep as much transparency as possible to the reader. Consider the whole network of blood

vessels, which are of different sort; e.g., in a more refined view, a heart chamber behaves differently than capillaries that compose the human cardiovascular system. This raises in fact a non-trivial challenge of semantic management. Moreover, such a mathematical modeling problem is actually approached with a multi-scale technique [Blanco et al. 2009], see Fig. 3. Lumped 0-D components (Fig. 3 on the left) are coupled to 1-D components (the arterial tree, see Fig. 3 on the right) and even 3-D components (small parts of an artery deserving closer attention; e.g., the region nearby an aneurism, see Fig. 3 on the top center). In a 1-D perspective, the lumped simplification does not hold anymore. The hypothesis in that case is that (***h*****$_4$**) "*the blood flows radially over the vessel [now, a pipe] as a viscous fluid*." (obeying (***m*****$_4$**) Poiseuille's law). In a 3-D perspective, the tension exerted on the vessel wall can be predicted precisely under the hypothesis that (***h*****$_5$**) "*the blood behaves as a viscous fluid*." (modeled by (***m*****$_5$**) the Navier-Stokes equations), see Table 1. In case a pathology is under investigation, modifications of many sorts are formulated as hypotheses to be tested through modeling and simulation. Fig. 3 highlights that scientists think over multiple hypotheses and models in an integrated way. Thus, linking hypotheses and models as (conceptual and logical data elements) to their workflow data elements at the physical layer allows for their semantic management, which could bring to the eScience architecture genuine benefits.

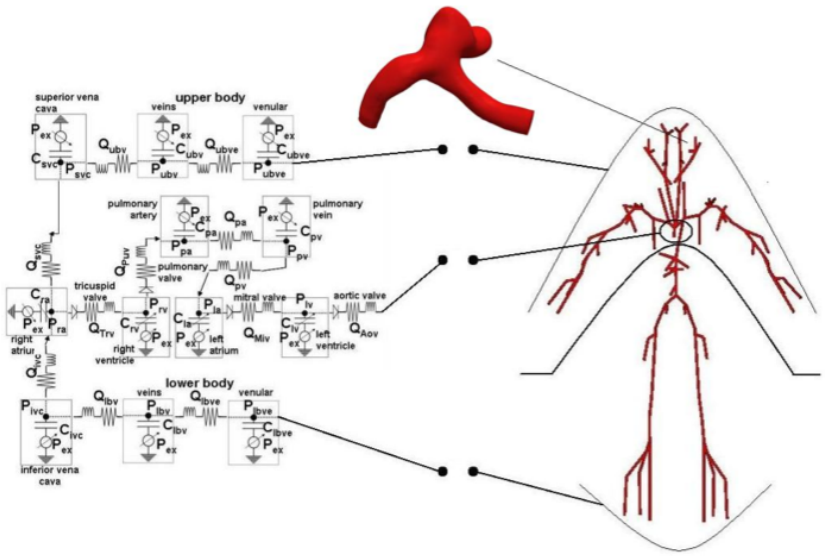

Fig. 3: Schematic diagram and scientific visualization rendered from models of the human cardiovascular system that embody different scientific hypotheses w.r.t. the behavior of blood vessels.

5. Conclusions

Modern science is confronted with a data deluge produced by new instruments and computer simulations of natural phenomena. In order to support scientists in making sense out of these data, eScience research has evolved by focusing on supporting the scientific experimental life-cycle *in-silico*. Scientific workflow systems allows scientists to design their experiments and to cycle through a validation-tuning experimental life-cycle.

In this paper we have proposed an extension to the experimental life-cycle for managing hypotheses and models. We argue that support to the *in-silico* scientific endeavour could go beyond scientific workflows and towards a complete scientific research life-cycle. One which includes the investigated phenomenon, the formulated scientific hypotheses and their related models, in addition to the experiments. This new vision of eScience support to the scientific research life-cycle has been materialized into an integrated three-layered architecture, comprising a conceptual, logical and physical layer, in which the physical layer has been already addressed by off-the-shelf scientific workflow systems. The integration fostered by the proposed architecture extends provenance information through the complete scientific research life-cycle, supporting collaboration, results interpretation and reproducibility.

There are many opportunities for future work. We are currently working on the management

issues regarding the scientific life-cycle in which hypotheses and models evolve, as well as in implementing data representations of both hypotheses and models using linked-data standards.

Acknowledgements

This work has been partially supported by CNPq (Conselho Nacional de Pesquisa), grants n°. 309502/2009-8, 382.489/2009-8 and 141838/2011-6.

References

Altintas I., Berkley C., Jaeger E., Jones, M. Ludascher B. and Kepler M. (2004) **"**An Extensible System for Design and Execution of Scientific Workflows", In: SSDBM'04**.**

Bavoil,L., Callahan, S.P., Crossno,P.J., Freire, J., Scheidegger, C.E., Silva, C.T. and Vo, T.H. (2005), *VisTrails: Enabling Interactive Multiple-View Visualizations.* In Proceedings of IEEE Visualization.

Blanco, P.J., Pivelloa, M.R., Urquizar, S., and Feijóo, R., (2009) *On the potentialities of 3d-1d coupled models in hemodynamics simulations.* Journal of Biomechanics, 42(7):19-930.

Bonner A. J. (1990), *Hypothetical Datalog: Complexity and Expressibility***.** Theoretical Computer Science 76, pp. 3-51, North-Holland.

Callahan A., Dumontier, M., Shah, N.H. (2011) *HyQue: Evaluating Hypotheses using Semantic Web Technologies*. Journal of Biomedical Semantics, 2(Suppl 2):S3.

Dean, J. Ghemawat, S. (2004), *MapReduce: Simplified data processing in large clusters*, Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, USA.

Hunter J. (2006), *Scientific Models – A User-oriented Approach to the Integration of Scientific Data and Digital Libraries***,** VALA, Melbourne, February**.**

Liang F., Takagi, S., Himeno, R., Liu, H. (2009) *Multiscale modeling of the human cardiovascular system with applications to aortic valvular and arterial stenoses.* Med Bio Eng Comput 47(7):743-55. DOI: 10.1007/s11517-009-0449-9.

M. Mattoso, C. Werner, G.H. Travassos, V. Braganholo, L. Murta, E. Ogasawara, D. Oliveira, S.M.S. da Cruz, and W. Martinho (2010). *Towards Supporting the Life Cycle of Large-scale Scientific Experiments.* Int Journal of Business Process Integration and Management, 5(1):79–92.

Oinn T., Greenwood M., Addis M. (2000), *Taverna: Lessons in Creating a Workflow Environment for the Life Sciences***.** Concurrency and Computation: Pract. Exper., 1-7.

Ogasawara, E., Dias, J., Oliveira, D., Porto, F., Valduriez, P., Mattoso, M. (2011), *An Algebraic Approach for Data-Centric Scientific Workflows,* VLDB, Seattle, Wa, USA.

Popper K. (2002)**,** *The logic of scientific discovery*. Routledge, 2nd ed.

Porto F., Tajmouati O. Silva V.F.V, Schulze,B., Ayres F.M. (2007), *QEF Supporting Complex Query Applications*. 7th Int'l Symposium on Cluster Computing and the Grid, Rio de Janeiro, Brazil, pp. 846-851.

Porto F., Spacappietra, S. (2011) *Data model for scientific models and hypotheses*. In "The evolution of Conceptual Modeling: From a historical perspective towards the future of Conceptual Modeling", LNCS, vol. 6520, pp 285-305. Springer, 1st ed..

Porto, F., Macedo, J. A. F., Tamargo, J. S., Zufferey, Y. W., Vidal, V. P., Spaccapietra, S. (2008). *Towards a Scientific Model Management System*. ER Workshops: 55-65.

Racunas S.A., Shah N.H., Albert I., Fedoroff N.V. (2004a) *Hybrow: a Prototype System for Computer-Aided Hypothesis Evaluation*, Bioinformatics, Vol.20, Suppl.1, pp. 257-264.

Sowa, John F. (2000), *Knowledge Representation: Logical, Philosophical, and Computational Foundations,* Brooks Cole Publishing Co., Pacific Grove, CA.