

Management of Scientific Experiments in Computational Modeling: Challenges and Perspectives

Regina Braga¹, Fernanda Campos¹, José Maria N. David¹, Jairo de Souza¹,
Leonardo Azevedo², Kate Revoredo²

¹Computer Science Department, Universidade Federal de Juiz de Fora, Juiz de Fora,
Brazil

²Computer Science Departament
Universidade do Rio de Janeiro, Rio de Janeiro, Brazil

{regina.braga,fernanda.campos,jose.david, jairo.souza}@ufjf.edu.br,
{azevedo, katerevoredo}@uniriotec.br

***Abstract.** Currently computer is essential to the success in conducting scientific research. In this context, e-Science appears as science performed with computer support aiming efficiency. The challenge, “Computational Modeling of artificial, naturals and socio-cultural complex systems and man-nature interaction” from SBC Great Challenges is strongly related to the e-Science context. The goal of this challenge is to create, evaluate, modify, compose, manage and exploit computer models in fields related to complex, artificial, natural, socio-cultural and human-nature systems. Technologies like semantic web service composition, data provenance, peer to peer networks and scientific software product line can be used as the basis for the specification and development of an e-Science infrastructure to handle challenges and solve problems. This paper discusses the main challenges involved in developing an e-Science infrastructure, presenting research goals for the next years.*

1. Introduction

The term Computational Science, created to contrast with Computer Science, has been used to designate models, algorithms and computational tools as a solution of complex systems of different natures [Roure et al, 2003]. Through these applications scientists, from other knowledge domains can investigate problems that could not recently be considered because of the high volume of data, the absence of analytic solutions, or the impracticability of studying them in laboratories. In this context, the term e-Science can be employed to describe the development of software services infrastructure capable of providing access to remote functionalities, distributed computational resources, information stored in dedicated databases, data, results and knowledge dissemination and sharing. The following challenge: Computational Modeling of artificial, naturals and socio-cultural complex systems and human-nature interaction is strongly related to e-Science context, considering that the goal is to create, evaluate, modify, compose, manage and explore computational models for domains related to artificial complex, natural and socio-cultural systems and the human-nature interaction.

There are several problems related to e-Science support. However, technologies such as peer-to-peer network, ontologies, composition and orchestration languages for semantic services, software product lines and data provenance management can be used as a base for the construction of an infrastructure to support e-Science. Despite these

technologies there are still some constraints due to the inability in solving problems associated to each technology and to benefit their integration, but they seem to be promised. We can highlight in this paper four challenges to support e-Science infrastructure:

- Creation of peer-to-peer scientific networks;
- Use and composition of semantic web services related to scientific context;
- Data provenance management and process execution;
- Creation and development of a scientific software product line.

We have been investigating for several years these challenges through the proposal of an infrastructure for e-Science [Silva et al., 2012] [Mendes et al., 2011] [Matos et al., 2009]. While the semantic web services support interoperability across platforms and operating systems, scientific networks need to allow researchers to establish collaboration networks, besides the support to web services composition for the scientific workflow specification. Ontologies support data provenance management as well as the integration between scientific networks and web services, mainly in experiment composition for scientific workflow. Afterwards, services and processes associate to scientific context will be able to take part in a Software Product Line aiming to support reuse of scientific applications for e-Science. The goal of this paper is to argue the main challenges related to the development of an infrastructure for e-Science support. The discussions will be addressed to detail research challenges related to e-Science context for the next years.

This paper is organized in the following sections. In Section 2, the concepts and technologies are presented. In section 3, the future researches are considered and argued in e-Science context. Finally, in section 4 some conclusions are discussed.

2. Requirements for an E-Science Infrastructure

Computational modeling involves a large set of algorithms and simulation techniques, data manipulation, data mining, among others, in which the model is a product of the research itself. These products can be interpreted as part of a computational process that filters, transforms, merges and generates data. This often requires cooperation between Computer Science scientists and others from different areas. Typically, in computational modeling there is some uncertainty about the model itself, since it involves a large number of parameters that must be explored and adjusted [Parastatidis, 2009]. Considering these questions and the "Computational Modeling of Complex Systems" challenge, the context of e-Science research to support the modeling of complex systems is challenging, especially when considering experiments related to bioinformatics [Matos et al., 2009], where the volume of data is huge and the need for cooperation among distributed applications is inherent.

E-Science can be defined as concepts related to the integration of computing in scientific research in various areas, considering that this computational support is of great importance to increase the efficiency of these studies [Roure et al., 2009]. Currently, computer systems have become important for scientific research, supporting all aspects of its life cycle. The term e-Science was introduced in the UK by Taylor [Matos et al, 2009], encapsulating technologies needed to support collaborative,

multidisciplinary research that has emerged in various fields of science. Taylor acknowledged the importance of using computational tools in collaborative, multidisciplinary scientific research, and used the term to encompass tools and technologies required to support such research. This use of computing resources in the development of research benefits the scientific communities, facilitating the sharing of data and computational services, and contribute to building a data infrastructure and a distributed scientific community.

In this context, the computing power has been widely exploited by scientists, for example, for the creation of sophisticated simulations related to climate and earthquakes, for extracting results from physical or astronomy scientific data, and for creating simulations and data analysis related to Biology. However, such applications also require an infrastructure that facilitates the use of a computing support apparatus. Often scientists in this field may not be able to use the available computing power efficiently. Thus, the area of e-Science is seeking to enhance and facilitate the use of this computation through an infrastructure that allows the design, reuse, annotation, validation and document sharing of artifacts generated by scientific research. Furthermore, all information must be managed efficiently by several processes, e.g., storage, retrieval and integration.

Thus, an infrastructure for e-Science raises several requirements that must be considered. Here are some relevant ones:

- **Storage:** the infrastructure must be able to store and process large volumes of data efficiently, regardless of their geographical location;
- **Transparency:** the user of the infrastructure should be able to discover, access and process data transparently, regardless of where these data are located;
- **Communities:** the infrastructure must be able to allow users to create, maintain and dispose communities, whether restricted or not. To do this, it is necessary to manage profiles creating rules to define permissions to the resources provided by communities;
- **Mobility:** the scientist should be able to access the available data from any device, including computers and mobile ones;
- **Workflow:** it must support the automation process, describing the scientific processes as clear as possible, and computationally processable, thus creating a fully automated workflow;
- **Provenance:** sufficient information should be stored at runtime, providing evidence of the concision of generated data, providing reuse of results and enabling the reproduction of scientific experiments;
- **Notifications:** Scientists should receive notifications about new available data and services according to their interests;
- **Decision Support:** The infrastructure must provide information and suggestions relevant to scientists according to their needs;
- **Scalability:** The infrastructure needs to support its own growth. The increase number of scientists, data and services should not affect or limit its use;

- Reuse of processes and services: The proposal must provide support and management to ensure the reuse of artifacts, tools and services in scientific workflows.

3. Computational Modeling and Software Engineering: Research Challenges and Opportunities

E-Science researches are gaining importance both in national and international scenario. However, there are still few Software Engineering groups that focus their researches on this field. Thus, there are some important challenges relating computational modeling, e-Science and Software Engineering that should be addressed.

Considering the challenges previously presented, and the researches that were developed by the authors in last years, discussions concerning some important research areas are presented. It is not our goal to broadly address all research areas related to e-Science, but rather focus on topics in which the group has expertise and believes that they can contribute for the advances in this area in the next years.

3.1. Peer-to-peer Scientific Network Creation

E-Science offers a promising vision of how information technology can help to improve the process of scientific research. However, currently, there is a gap between the efforts already made and the vision of e-Science. A higher degree of usefulness and automation should be considered, suggesting opportunities for collaboration and computing on a global scale [Parastatidis, 2009].

The traditional client/server model used by the web can bring many problems when applied to e-Science. Performance of centralized servers can decrease due to the large volume of data provided by scientific experiments and the increased demand for processing power [Lua et al., 2005]. Therefore, causing some service access problems to scientists during scientific experimentation. These problems can be solved by the use of clustering techniques for multiple servers. However, besides not being a simple solution to be implemented, the clustering is also a very costly one, which very often constrains its implementation.

Therefore, decentralization has proved to be a solution to the presented problems, but it brings other challenges [Roure et al., 2003] and [Roure et al., 2001]. The peer to peer networks in turn, uses resources from multiple computers geographically distributed and are a good alternative for working with completely decentralized distributed architectures.

The peer to peer network has proved efficiency for data sharing. Considering that scientific information is naturally distributed and has free access, a peer to peer network appears as a outstand solution. By using peer to peer networks in conjunction with e-Science, we have an access point for scientists to access the knowledge base dispersed among scientists connected to the network. In addition, we have the possibility to connect existing scientific applications and use the features available on the network, using its low cost of ownership and maintenance when compared to a distributed system.

Recently, peer to peer networks have also been used successfully for interconnecting large distributed and heterogeneous scientific databases, enabling the

exchange of research on complex data structures [Loser et al., 2003]. A problem with this approach is that most of these structures are too complex for scientists, and does not have a semantic data description, thus making it more difficult for understanding. One promising proposal is to add a semantic search for data and applications layer in scientific peer to peer networks, creating semantic communities for scientist groups considering their domain, expertise and interests. Considering also that there are several scientific communities with a large volume of data spread in a completely decentralized environment, a proposal is to use peer to peer networks in conjunction with an e-Science infrastructure, facilitating and enhancing the access of scientists to the large volume of scientific distributed data. But the peer to peer networks tend to have a high volume of communication demand when the flow of scientific work increases. As the scientific research focuses on small communities with few connections between them, you can limit communication between network nodes considering only those working in this scientific community, thus reducing the communication between nodes and increasing the efficiency of the search on the network.

Experiments simulations show that the selection of nodes based on expertise to create a group increased the performance of peer to peer networks. Considering this context, the semantic web may help in the management of the specialties of the scientists, facilitating the specification of research groups composed by nodes in the network [Berners-Lee et al., 2001]. Using ontologies to describe the groups and their specialties, we can discover relationships between relevant distributed researches groups scattered around the network. The advantage of this approach is that the process and results of a given experiment will not be distributed to a number of nodes randomly, but only for those that may have relevant information to the specialization of that research.

The Semantic Web aims to solve the problem of information complexity, providing support for advanced means of representing and processing information. The peer to peer networks can address the complexity of the system, enabling a flexible and decentralized structure to store and process information. Making the conjunction of these two technologies to solve some of the problems offered by e-Science, we can achieve an efficient, inexpensive and sometimes redundant infrastructure to be used by scientists.

Analysing the requirements presented in section 2, semantic peer to peer network can help to fulfill the presented requirements, specially the following (i) communities, once we can create restricted groups related to a given scientific domain; (ii) workflow, considering that we can compose services in restricted scientific peer to peer network; (iii) notifications, since peer to peer architecture provides this mechanism; and (iv) scalability, providing mechanism to the scientific network expansion.

3.2. Data Provenance Management and Process Execution

Considering the scientific context, one can argue that the benefits offered by computational tools imply in a new and complex challenges in scientific research scenario. Thus, it is possible to point out some important questions. A key aspect is related to the reuse of generated knowledge. Scientific experiments computationally processed are subject to frequent updates, considering, for example, more refined view

of the researchers or the changes in experiment stages and their tasks, which may become of greater or lesser relevance when the model is refined. In addition, databases can be updated periodically, which leads to deprecated results that were previously obtained. Therefore, knowledge reuse can mean saving time and resources to scientific community. Another relevant aspect refers to results of an experiment computationally processed, which may have little utility if scientists are unable to judge the suitability of the analyzed problem or identify the source of results. In a scenario of scientific research, we can consider that part of the data meaning is due to understanding of the generative process.

Data provenance is the description of the origins of a data item and the process by which it was produced. The data provenance helps to form a vision of quality and validity about the information produced today in the context of a scientific experiment [Buneman et al., 2001]. This information aims to describe previous data that may have been generated during the experiment execution and present the transformation processes to which these data was submitted. For scientists, the provenance of scientific experiments can indicate how the results were derived, which parameters influenced the derivation process and which databases were used as input for the experiment processing.

However, in order to obtain the benefits of provenance data, it is essential that the information related to computer simulation that represents the scientific experiment can be captured, modeled and persisted properly for later use. In this context, we can notice that the management of data provenance has received considerable treatment by the scientific community. Among the problems studied by the provenance community it is important to mention the lack of agreement about the scope of information to be captured in a provenance model. The Open Provenance Model (OPM) is a generic and comprehensive representation of the data provenance beyond the scope of scientific experiments [Moureau et al., 2008] and it has been used in some successful projects.

The OPM model can be used in a context where information originally stored in heterogeneous repositories and scientific workflows modeled in different systems lack mechanisms that contribute to a better interoperability between the data generated and processes orchestrated within collaborative research projects. Thus, it is necessary to provide a layer to make such interoperability possible. It includes the collection and management of data provenance and applications in a context of scientific experiments processed through computer simulations in collaborative research environments, geographically dispersed and interconnected through a computational grid.

This provenance layer of an e-Science infrastructure must be based on some important requirements, including (i) the layer must be independent of control flow mechanisms and data formats implemented by existing SWfMS (Scientific Workflow Management Systems), (ii) it must be applicable in a wide range of scientific experiments, including those running on heterogeneous environments - from a variety of hardware platforms and operating systems - and dispersed in a computational grid, (iii) it must assess the impact of collection, persistence and query provenance metadata on the performance of scientific experiments, (iv) it has to use semantic web technology - RDF, OWL ontologies, inference engines - for representation and query the provenance metadata, (v) it can use a collection and management of provenance data at varying levels of abstraction, allowing the parameterization of the provenance graph in

accordance with the interests of ongoing research providing results with greater or lesser extent and detail considering the same graph provenance.

The use of semantic web technologies is important in this context, considering ontologies, and the OPM model. The use of ontologies allows the construction of new specialized OPM profiles in order to incorporate new attributes into the provenance semantic graph. These specialized profiles extend the power of semantic processing in scientific workflow domain, extracting information inferred from semantic queries.

Research on data provenance is an open issue in e-Science context. Important research questions are:

- a) How to construct an instrumentation mechanism, in order to optimize performance of metadata provenance collection and persistence? One of our hypotheses is to distribute provenance data storage on different nodes of the computational grid.
- b) What are alternative technologies for semantic query provenance? It can be argued that SPARQL language, although considered a standard by W3C, imposes limitation on its use due to its complexity. One possibility is to use a graphical representation of lineage graph to support query definition from a visual and intuitive interface.
- c) OPM model is evolving considering data provenance challenges [Moureau et al., 2008]. It is important that provenance research investigates also these challenges.

From the e-Science requirements presented previously, we can highlight the provenance requirement, as that which can benefit directly from a provenance management system proposed in this article. The decision support requirement is also fulfilled by the provenance management system, once it can provide strategical information for users, also including an efficient storage system.

3.3. Semantic Web Services Use and Composition Related to Scientific Context

E-Science activities are accompanied by a proliferation of data and tools. This brings new challenges, for example, how to understand and organize these resources, how to share and reuse successful experiments (tools and data), and how to provide interoperability among data and tools from different locations and used by people with different profiles.

Web services have been presented as a viable solution for remote execution of functionalities. In part this is due to some properties of Web Services such as independence of operating systems and programming languages, interoperability, ubiquity and the possibility of developing systems loosely coupled. However, in the e-Science context, to connect applications between geographically dispersed research groups it is necessary to encapsulate these scientific applications as Web Services, to make available a more semantic description of their functionality, so that research groups can find remote services that best fit their needs, but also fulfill non-functional requirements. In addition, we may need to compose these Web services in a scientific workflow, where information from different research groups can be processed by different applications in order to obtain a final result [Medjahed and Bouguettaya, 2011].

Workflow technology provides room for the resolution of problems of scientific nature. It facilitates the creation and execution of experiments using a large amount of

available data and services. Scientific workflows are being increasingly adopted as a means to specify and coordinate the implementation of multidisciplinary experiments. They allow the representation and execution of tasks using heterogeneous tools and data. In this context, the use of semantic tools to support searching for resources to compose scientific workflows is a key issue, considering that the choice of the exact resources for each task that makes up a process is not straightforward.

Given to the growth of the Web both in size and in diversity, there is an increasing need to automate some aspects related to Web Services, such as discovery, execution, automatic selection and composition. The use of Semantic Web technologies, such as semantic Web Services and ontologies, can support in the construction of tools that enable this automation. On the other hand, the creation of scientific workflows is not a straightforward task. It is difficult to formally define scientific experiments. The task of creating scientific workflows with available Web Services becomes even more complex because most scientists are not computer scientists, and therefore it is not easy to use computer resources to carry out their scientific experiments.

One of the challenges in this area involves the use of Semantic Web as a way to enhance the discovery and composition of Web services for the development of scientific workflows, including an architecture that helps the scientific community to define and execute scientific experiments through the use of semantic Web Services organized in a scientific workflow considering a peer to peer network. It involves research on semantic repositories, ontologies, Web Services composition and scientific workflows.

Considering this scenario, we can enumerate some research opportunities in this area, such as (i) management of distributed and heterogeneous scientific repositories, (ii) searching and retrieval of scientific semantic Web Services, (iii) composition modeling, including semantic and syntactic compatibility, and (iv) ontology matching, considering semantic Web Services from different domains. We believe that, despite of the discussions and some results, much more effort needs to be done. The requirements that are more connected to the problems and possible solutions described above are workflow, transparency, mobility and scalability.

3.4. Software Product Line Development and Management of Artifacts Related to Scientific Domain

Artifacts reusability and maintainability of a family of software products can be enhanced through the Software Product Line (SPL) approach [Clements and Northrop, 2002]. In scientific context it could be advantageous to develop a product family in order to support scientific experiments. Most of time scientific workflow users do not have enough knowledge about software development and do not start from scratch. As a result, they can develop an application based on an existing workflow and do adjustments according to their needs.

Considering the benefits of SPL in application development context with a high degree of similarity, our research group is working to associate domain ontologies and feature models in order to support the development of a scientific SPL. Features models can be used as a way to represent variabilities between products. Although these models offer an easy way to understand concept relationships, they do not describe the semantics rules offered by ontologies. They can also enhance variabilities specification,

supplying additional information for the domain for which a feature model is built. Another challenge to achieve is to enhance reusability through the development process and/or the execution of scientific applications. Through the connection between these models a scientific workflow could be reused with the activities that are accomplished according to the chosen domain and user's requirements. SPL could support scientists' choices as well as the definition of a workflow process, according to their needs [Costa et al., 2012].

However, some challenges still need to be addressed in order to enhance SPL usage in scientific experiments, such as (i) services search and recover that is more adequate to the application to be developed, (ii) services selection to compose scientific workflows, and (iii) services development related to workflows interoperability. Considering the requirements presented in section 2, the transparency can benefit directly from a SPL approach, since it can help scientist in scientific application/workflow development, hiding implementation issues. A SPL can also help on decision support, considering that it may bring strategical information, based on the artifact repository of the SPL.

4. Related Works

In e-Science literature, there are some initiatives related this research. In UK there are several projects related to e-Science, covering areas such as astronomy and earth science [International Virtual Observatory Alliance, 2012]. High-performance simulations, computational steering, and remote visualization to advance the state of the art are also used in these projects [GENIE Project 2012]. In bioinformatics, researchers and pharmaceutical companies are attempting to use e-Science technologies to reduce data to information and information to knowledge [myGRID project, 2012], [e-Family Project, 2012].

The myGrid e-Science project is researching high-level middleware to support personalized in silico experiments in biology [MyGRID project, 2012]. The emphasis is on data-intensive experiments that combine the use of applications and database queries.

Another important project, CombeChem, has the goal of creating a Smart Laboratory for chemistry, using technologies for automation, semantics, and Grid computing [The CombeChem project, 2012]. A key challenge of the project is the fact that large volumes of new chemical data are being created by new high-throughput technologies, such as combinatorial chemistry, in which large numbers of new chemical compounds are synthesized simultaneously. The need for assistance in organizing, annotating, and searching this data is becoming relevant.

De Roure, Jennings, and Shadbolt (2001) introduced the notion of the Semantic Grid, which advocated "the application of semantic Web technologies both on and in the Grid". They identified a need for maximum reuse of software, services, information, and knowledge. They consider the e-Science requirements as a spectrum, with one end characterized by automation, virtual organizations of services, and the digital world, and at the other end they are characterized by interaction, virtual organizations of people, and the physical world.

4. Conclusions

Computational resources are becoming increasingly important in the life cycle of a scientific research. Thus, the amount of scientific data computationally generated are increasing and the computational power enables the execution of these experiments that when not automated, could hardly be executed. However, we have the growing on the volume of data in the scientific context, that provides some challenges on the storage, processing and searching of scientific data.

This paper aimed to discuss the development of scientific research in a distributed context. We considered use of technologies such as the peer to peer networks that provide a low cost, but great potential for data sharing, the Semantic Web, which provides a rich semantic description of the available information, Data Provenance that allows the management of historical data, Product Lines that provides and infrastructure to facilitate the development and reuse of scientific applications related to a given domain and Web Services, it can allow in an easy way the composition of scientific applications in order to execute an experiment.

The development of computational scientific experiments is a complex process. There are different components (applications) which must be used together to perform one experiment. Therefore, research in this area should consider the provision of an infrastructure that supports the inclusion of new applications, which can interact with others, creating a framework that provides adequate functionality and data to scientists. We consider that the proposals outlined in this paper can contribute in e-Science research. We can highlight some important aspects:

- Mobility and accessibility improvement offered by the use of semantic web in a distributed environment;
- Creation and management of new scientific applications facilitated by the use of Web Services and the possibility of their composition in scientific experiments;
- Easy creation and deployment of new applications by scientists through the use of facilities provided by the semantic web, Web Services and peer to peer networks;
- Communication between applications increased, improving the functionality and data availability considering data provenance and composition of scientific applications;
- Grouping support of scientific network nodes according to their specialties and interests, providing greater accuracy and research collaboration according to its semantic content;
- Semantic search and data storage resources improvement provided by the use of semantic web and data provenance.

Acknowledgment

This work is being accomplished in the context of the “Infrastructure for the Management of Scientific Experiments in Computational Modeling” project, granted by CNPq, no 559998/2010-4.

References

- Berners-Lee, T., Hendler, J. and Lassila, O. (2001) "The Semantic Web", *Scientific American*, May, 284(5):34–43.
- Buneman, P., Khanna, S. and Tan, W. C. (2001) "Why and Where: A Characterization of Data Provenance". In: *ICDT '01: Proceedings of the 8th International Conference on Database Theory*, Springer-Verlag: London, UK, p. 316–330.
- Clements, P. and Northrop, L. (2002) "Software Product Lines: Practices and Patterns", Addison-Wesley Publishing Company, Boston.
- CombeChem project (www.CombeChem.org), accessed in dec. 2012.
- Costa, G., Braga, R., David, J. M. N. and Campos, F. (2012) "Connecting Feature Models and Ontologies in Software Product Lines", In: *14 International Conference on Enterprise Information Systems*, Wroclaw.
- De Roure, D., Jennings, N. and Shadbolt, N. (2001) "Research Agenda for the Semantic Grid: A Future e-Science Infrastructure", *Grid Computing - Making the Global Infrastructure a Reality*, ISBN 0470853190, Pages 437-470.
- De Roure, D., Jennings, N. and Shadbolt, N. (2003) "The Semantic Grid: A Future e-Science Infrastructure | Concurrency and Computation: Practice and Experience", Volume 15, Issue 11.
- e-Family Project (www.sanger.ac.uk/xml/eFamily.xsd), accessed in dec. 2012.
- GENIE Project (www.genie.ac.uk), accessed in dec. 2012.
- International Virtual Observatory Alliance (www.ivoa.net), accessed in dec. 2012.
- Löser, A., Wolpers, M., Siberski, W. and Nejdil, W. (2003) "Efficient data store discovery in a scientific P2P network", In: *Proc. of the WS on Semantic Web Technologies for Searching and Retrieving Scientific Data*, CEUR WS 83.
- Lua, E. K., Crowcroft, J., Pias, M., Sharma, R., Lim and S. (2005) "A Survey and Comparison of Peer-to-Peer Overlay Network Schemes", *IEEE Communications Surveys and Tutorials*, Volume 7, Pages 72-93.
- Matos, E. E., Campos, F., Braga, R. and Palazzi, D. (2009) "CelOWS: an ontology based framework for the provision of semantic web services related to biological models", *Journal of Biomedical Informatics*, v. 43, p. 125-136.
- Medjahed, B. and Bouguettaya, A. (2011) "Service Composition for the Semantic Web", Springer, ISBN 978-1-4419-8464-7.
- Mendes, L. F., Silva L., Matos, E., Braga R. and Campos, F. (2011) "SASAgent: an Agent Based Architecture for Search, Retrieval and Composition of Scientific Models"; *Journal of Computers in Biology and Medicine* ISSN 0010-4825.
- Moreau, L., Freire, J., Futrelle, J., McGrath, R. E., Myers, J. and Paulson, P. (2008) "The Open Provenance Model: An Overview", pp. 323–326.
- myGrid Project (www.mygrid.org), accessed in dec. 2012.

Parastatidis, S. (2009) "A Platform for All That We Know: Creating a Knowledge-Driven Research Infrastructure", In Hey, T., Tansley, S., Tolle, K. The Fourth Paradigm: Data Intensive Scientific Discovery. p. 165-172.

Silva, L. M., Braga, R. and Campos, F. (2012) "Composer-Science: A semantic service based framework for workflow composition in e-Science projects", Information Sciences, v. 186, p. 186-208.