Proceedings of the 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE), Larnaca, Cyprus, 11-13 November 2012

Open Source Workflow Systems in Life Sciences Informatics

Achilleos K. G., Kannas C. C., Nicolaou C. A., Pattichis C. S.
Department of Computer Science
University of Cyprus
Nicosia, Cyprus
{cnicolaou, pattichi}@ucy.ac.cy

Promponas V. J.
Department of Biological Sciences
University of Cyprus
Nicosia, Cyprus

***Abstract*—A simple yet powerful programming tool enabling *in silico* experimentation, *end-to-end* data management through web services as well as use of grid and cloud processing power is scientific workflows. This technology is receiving considerable interest in recent years primarily due to its ability to promote and support scientific collaboration among large distributed research teams. The paper reviews the Scientific Workflows Management Systems (SWMS) field and investigates in detail popular open source workflow systems used commonly in life sciences informatics. Emphasis is placed on features which make these systems attractive for scientific use, e.g. user friendliness, use of distributed resources, reusability, provenance, collaboration, data integration, etc. Our conclusions indicate that although SWMS, including open source ones, have several open issues, their unique features and strong momentum clearly suggest that it is only a matter of time before they are adopted in even more scientific fields.**

Keywords-scientific workflow; scientific workflow management system; in silico experiment

I. INTRODUCTION

Nowadays scientists work in e-science environments and carry out *in silico* experiments. In other words scientists use environments that support global collaboration, involve multidisciplinary science and utilize modern technology infrastructure [1] to carry out their experiments *in silico*. A powerful approach, with proven capabilities to facilitate the design process of computational experiments is based on scientific workflows (SW). This approach enables scientists to plug together problem solving computational components [2] and implement complex *in silico* experiments such as the analysis of large datasets that arise from sensors or computer simulations, and, the design and execution of complicated algorithms with numerous computationally intensive steps. SWMS can potentially accelerate scientific discovery by incorporating data management, analysis, simulation, and visualization tools. They provide an interactive visual interface that facilitates the design, execution and management of workflows. Moreover, SWMS enable remote access as well as data and services sharing, making possible collaborations among geographically distributed researchers.

SWMS have quickly found application in several, diverse scientific domains. This domain independence is mainly owed to the abstraction that characterizes the workflow paradigm. Fig. 1 illustrates some of the main application domains of SWMS. As can be seen, life science related disciplines are heavily represented. In fact, it is the case that many current workflow systems began their development from a life science related project.

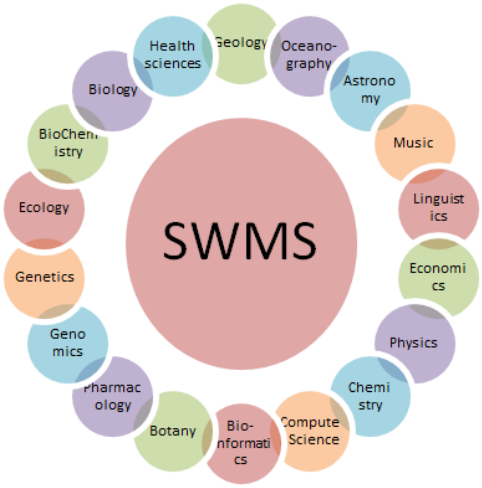

Figure 1. Application domains of SWMS

Recent advances in SWMS technology do not yet match the expectations of scientists [3]. However they are a step towards a future where we can imagine a doctor preparing a checkup prognosis and therapy workflow of a patient based on complicated DNA analysis, statistical prediction models, image analysis algorithms, inference rules engines and custom drug selection/design all from a tablet pc only to be executed somewhere in the cloud.

The remainder of this review paper is structured as follows: section II briefly describes scientific workflows and explains the reasons for their recent popularity. Section III provides a more detailed description of the technology while section IV presents selected widely used scientific workflow systems. The paper continues to section V with a critical review based on the experiences obtained from the implementation of a computational experiment in two representative systems and concludes in section VI.

II. WHY SCIENTIFIC WORKFLOWS

Traditionally, many scientists have been using batch files, shell scripts and programs written in general-purpose scripting languages (e.g., Perl, Python) to automate their tool-integration tasks [3]. This approach provides high flexibility, and is

therefore appealing, to expert users but makes it difficult for the average user to implement scientific tasks requiring the integration of multiple computational components and data resources. Scientific workflows provide a promising alternative to all researchers facing the above problem because of several inherent advantages. Two main advantages of the SW approach are visual representation of the task flow and visual channeling of data as opposed to lines of code directing the flow in the case of scripts. Provenance information, which is very important for the reproducibility of the experiments as well as for backtracking and resolution of errors, is an additional characteristic of workflows not present in scripting tools. Reusability and transparency is easily achieved by the reuse of a workflow or the use of a workflow inside a workflow. Finally complex implementation details such as parallelism, pipelining and High Performance Computing (HPC) are handled transparently by SWMS systems in order to achieve maximum efficiency for execution time.

Fundamentally, a scientific workflow is a tool that automates the execution of an experiment. As such it can offer multiple benefits for all the phases of an experiment's lifecycle. During the composition phase, a repository of tried and tested workflows is available to the scientists to choose from. During the execution phase, as experimenting is by definition a repeatable process, workflows can relieve the scientists of repetitive tasks but at the same time keep track of all the intermediary steps and data. These traces can be used at a later stage to enable the reproducibility of the experiment. Provenance information is also useful during the analysis phase to assess the evolution of the research effort, trace the origin of an error or go back to a previous stage and change the direction of investigation. Visualization tools are provided for this phase as well for assisting in the evaluation of the results [4].

Scientific workflows can also serve as a tool for end-to-end scientific data management by enabling scientists to cope with big data produced through various scientific processes. Grid technologies allow workflows to implement parallel executions enabling large-scale data processing. In this case, workflows are used as a parallel programming model for data-parallel applications. Web services allow ease of access to local and distributed data sources as well as data aggregation from highly heterogeneous environments. Even HPC technology can be made available to scientists who may have limited or no computing resources. Finally, collaboration between scientists is encouraged and achieved both within and across disciplines. Implemented similarly to the trend of social networks, scientists share workflows and their corresponding services. All of the above can optimize the implementation of experiments in a transparent way for the domain scientist.

Currently over 50 different representatives of SWMS exist [5]. The most popular open source SWMS's in life science scientific literature are Taverna [6], [7], [8], and KNIME [9], [10]. Galaxy [11]-[13] is a more recent web based SWMS dedicated to biomedical research that is increasingly gaining popularity. Pipeline Pilot [14] and InforSense KDE [15] are commercial software products also widely used in the industry.

III. SCIENTIFIC WORKFLOW TECHNOLOGY

A. Scientific Workflow Paradigm

A workflow is a general, widely used term used to describe the actions that need to be taken in order to complete a complex task. An abstract scientific workflow is represented as a directed graph where each node represents a step implemented by a software component. This component can be either the execution of a local program or a remote web service (e.g. a query to a database). The edges of the graph represent either data flow or execution dependencies between nodes [30]. The links coordinate the inputs and outputs of the individual steps, forming the data flow. Control flow links occur when two tasks have no data dependencies and therefore the order must be explicitly defined.

In Fig. 2 a sample workflow, designed using the KNIME [9], [10] platform, is depicted. The sample workflow reads a file containing molecules, converts them to an internal structure so that its descriptors can be calculated and consequently writes the results in a file and at the same time enables the visual examination of the molecules. Each step is represented by a node which is clearly named. The links denote the flow of the data from one node to the next. The order of execution is determined by the data dependencies.

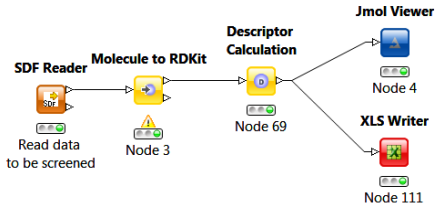

Figure 2. Example workflow in KNIME

Usually, real life scientific workflows are more complex with calls to more services, usage of shim services to convert between inputs and additional parameters for sequence of execution, looping and error handling.

Abstract workflows are sometimes described using special languages or XML schemas e.g. BPEL [16], [17] in the Trident system, DAG [18] in Pegasus, t2flow [19] in Taverna or even simple database values as in Galaxy [20]. Once the abstract workflows are translated into machine readable language they can be fed into workflow execution engines.

B. Types and Subcategories

Flow control can be considered as the most important classification characteristic of scientific workflows. A workflow is either data-flow or control-flow oriented. In control-driven workflows the connections between the tasks represent a transfer of control from one task to the next one. In data-driven workflows connections represent the flow of data from one task to the next one. The workflow representation is centered on data products. As mentioned in [2] most of the current scientific workflows are data-flow oriented as opposed to their predecessors and business workflows which are control-flow. According to [21], the reason is that data-flow modeling is the natural way of composing scientific

workflows, because they often comprise numerous data transformation steps.

Another important distinguishing feature of workflows is pipeline parallel processing. A pipeline consists of a collection of steps. Parallelism is achieved by executing these steps simultaneously on different input data sets. The tasks are executed in separate threads, processing input immediately and not waiting for the previous task to complete. The drawback is that pipelined workflows are harder to restart in the case of unforeseen events as the current state of the executed workflow is harder to describe and restore [4].

C. Scientific Workflow Management Systems

In theory a SWMS is a combination of workflow modeling components using an abstract language and a workflow enacting component empowered by an execution engine. In practice a SWMS enables a user to create and monitor the execution of a workflow by providing the necessary infrastructure. The modeling component enables the user to design, store and reuse workflow models while the enacting component invokes, executes and monitors workflow instances [5] deploying them either on a local desktop computer, a web server or a distributed computing environment. Embedded in the workflow design, is the order of the tasks to be executed. This architecture, known as centralized execution architecture, is applied in the Trident SWMS [22]. Other systems follow a less strict, decentralized architecture. For example, in Taverna 2 [6], each processor independently starts its own execution as soon as the input data are available. This allows for inter-processor parallelism as the tasks are executed in separate threads.

D. Scientific Workflow Collaboration

As previously mentioned, one of the main advantages of these tools is their ability to promote scientific collaboration through sharing of workflows. An example of such initiatives is myExperiment [23], a social networking site for researchers providing a Virtual Research Environment (VRE) designed for users to share, discover and reuse workflows [19].

Experts stress that workflows "encapsulate scientific intellectual property". As such they must be stored, organized and easily retrieved. myExperiment aims to be an online scientific workflow repository for organizing, sharing and discovering analogous to online research paper management applications. Researchers can publish the workflow to be easily accessed by interested scientists through myExperiment. Furthermore, users can tag and comment on workflows, and, create and join groups and exchange messages. Initially, myExperiment was built as part of the myGrid and Taverna projects for supporting bioinformaticians but is now used by a wider range of disciplines and supports different types of workflows. Currently myExperiment has over 5000 members, 250 groups, and 2000 workflows from Taverna users but also KNIME, Kepler, Pipeline Pilot and, more recently, Galaxy.

IV. OPEN SCIENTIFIC WORKFLOW MANAGEMENT SYSTEMS – SELECTED SYSTEMS IN LIFE SCIENCES INFORMATICS

The field of SWMS has been receiving considerable interest in recent years. Consequently, a number of implementations have been reported and several reviews of such systems have been published. This section provides an updated concise review of the most popular open source SWMS used in life science informatics research, in order to present the current state of the art in the field. Table 2.1 presents a more comprehensive snapshot of the main representatives of workflow management tools and their main characteristics. The interested reader can also find excellent reviews on the topic in Barker and van Hemert [2], Curcin and M. Ghanem [24], McPhillips et al.[3], C. Goble et al. [5], Ludascher et al. [4] and [25], and. Deelman et al. [26], and Sonntag et al. [22].

A. Taverna

Taverna is an open-source, grid-aware workflow management system [5]-[7]. It has found wide application in the bioinformatics, chemistry, data- and text-mining and astronomy communities although the system is domain independent. It is comprised of the Taverna Workbench graphical workflow authoring client, a workflow representation language, and an enactment engine. Taverna is implemented as a service-oriented architecture, based on web service standards. From the advent of its design Taverna was an application that applied web services technology to workflow design. That meant that tools created using different programming languages (e.g. Java, PERL, Python, etc) or platforms (Unix, Windows, etc) could now be accessed via a web service interface eliminating any need for integration. The same applied to the databases available on the web. As a result, researchers could design and execute a pipeline of web services, with little programming knowledge. Its architecture supports parallelism, both intra-process and inter-process, asynchronous service support and separation of data and process spaces to support scaling to arbitrary data volumes.

A vital component of Taverna's open architecture is the plug-in functionality. Various plugins have been developed for accessing online bio-catalogues or for integrating chemo-informatics processing services. Provenance also plays an integral part in Taverna, allowing users to capture and inspect details such as who conducted the experiment, what services were used, and what results were produced. An additional strong feature of Taverna is workflow sharing. The users have direct access to the myExperiment social collaboration site where they can upload or download workflows as needed.

B. KNIME

KNIME (Konstanz Information Miner) is a modular environment that supports operations such as data integration from various sources, processing, modeling, analysing and mining, as well as parallel execution [11], [12]. KNIME is primarily used in pharmaceutical research with some applications reported in other areas like customer resource

TABLE I. LIST OF SCIENTIFIC WORKFLOW APPLICATIONS

	List of Scientific Workflow Applications			
	Application	***URL***	***Techology/ Paradigm***	***Scientific field***
OPEN SOURCE	Taverna [6]-[8]	http://www.taverna.org.uk	Java, graphical interface, local instance, server based, grid interface	Bioinformatics, Chemistry, Astronomy, Data and Text mining, Music,
	Galaxy [11]-[13]	http://galaxy.psu.edu	Python, graphical interface, web based, grid or cloud instance	Life Sciences, Bioinformatics
	Pegasus [18]	http://pegasus.isi.edu/	Java, grid interface	Bioinformatics, Astronomy, Botany, Chemistry, Physics, Ocean science, Neuroscience, Limnology, Genome analysis, Earthquake science, Climate modeling, Computer science, Helioseismology
	Triana [27]	http://www.trianacode.org	Java, graphical interface, grid interface	Signal, Text and Image processing
	Kepler [28]	https://kepler-project.org	Java, graphical interface, grid and web services extensions	Ecology, Geology, Chemistry
	KNIME [9][10]	http://www.knime.org	Java based, graphical interface, local or web instance, server-based	Life Sciences, Chemo- and Bioinformatics, Data Analysis
COMMERCIAL	DiscoveryNet [15] Inforsence, IDBS	http://www.idbs.com/		Life sciences, Healthcare, Financial services, Sales & Marketing analytics, Environmental Monitoring, Geo-hazard modeling
	Pipeline Pilot[14]	http://accelrys.com/products/pipeline-pilot/		Biology, Chemistry, Material Science
	Microsoft Trident [29]	http://research.microsoft.com/en-us/collaboration/tools/trident.aspx		Oceanography, Astronomy

management and data analysis (CRM), business intelligence and financial data analysis. It is an open-source platform free for nonprofit and academic use. It is available as a local desktop application but additional features such as user authentication, web services integration, web browser interface, remote server or cluster execution, server execution are available in (and restricted to) the professional package.

The platform enables the user to visually assemble and execute data pipelines providing an interactive view of the results. KNIME pipelines consist of modular independent components that combine different projects in a single pipeline. At the same time its expandable architecture enables the easy integration of newly developed tools.

One highlight of KNIME's latest additions is the ability to support PMML[30]. The Predictive Model Markup Language (PMML) is an XML-based markup language that enables applications to define models related to predictive analytics and data mining and to share those models between PMML-compliant applications. As a result a model developed by KNIME can be exported and then used in another data mining engine. Another characteristic is the addition of database ports that are JDBC-compliant that work directly in the database enabling even preview of the actual data inside the database tables[30]. JDBC is a Java-based data access technology that provides methods for querying and updating data in a database.

Although written in Java, KNIME, permits running Python, Perl and other code fragments through the use of special scripting nodes. This is extremely useful as a lot of scientific work is currently under the form of Python or Perl scripts.

KNIME functionality is enriched by integrating functionality of different data analysis open source projects for machine learning and data mining, for statistical computations and visualizations as well as many chemoinformatics plugins.

C. *Galaxy*

Galaxy is a web-based platform for data intensive biomedical research [16]-[18]. It provides a framework for integrating computational tools and an environment for interactive data analysis, reuse and sharing. As stated in [16], [18] the primary design considerations of Galaxy were accessibility, reproducibility and transparency. Galaxy is accessible to scientists with no programming knowledge through the use of Galaxy tools. It produces reproducible computational analysis results by generating metadata for each analysis step through the automated production of Galaxy History items. It also promotes transparency by enabling the sharing of data, tools, workflows, results and report documents.

A structured well-defined interface allows the wrapping of nearly any tool that can be run from the command-line into a Galaxy tool. The platform is open source and has been designed specifically to meet the needs of bioinformaticians supporting sequence manipulation with built in libraries. It does not support any control flow operations or remote services. Additionally it does not use a workflow language but rather a relational database. The Galaxy workflow system allows for analysis using multiple tools incorporated to the system which may be built and run or extracted from past runs, and rerun.

Pages are a feature unique to Galaxy. They are online documents used to describe the analysis performed but also to provide links to the Galaxy objects that were used in the analysis, i.e. Histories, Workflows, Datasets. This enables the reader of the document to have direct access to the dataset used, to import the workflow and reproduce the experiment himself. It also makes it even easier for another scientist to continue and build upon reported previous work.

A recent Taverna-Galaxy integration allows the generation of Galaxy tools from Taverna 2 workflows [20]. The tools can then be installed in a Galaxy server and become part of a Galaxy pipeline. More over Galaxy workflows can be directly shared through the myExperiment site. Galaxy can also be instantiated on cloud computing infrastructures and interfaced with grid clusters [31].

V. CRITICAL REVIEW OF SWMS

In order to assess progress in the SWMS field and be able to evaluate what workflow technology currently offers, how we can benefit from it, how it can be improved and what difficulties arise during use, the authors implemented a complex computational experiment described in detail in [32]. The two systems selected were KNIME, due to its appealing interface and wide range of plugin tools, and Galaxy for its online nature. The scientific workflow designed addressed the needs of an *in silico* virtual screening (VS) experiment from the life sciences field, specifically, the chemoprevention domain.

This task involved the preparation of appropriate nodes/tools for each of the SWMS, the implementation and execution of the workflows and the analysis and presentation of the results obtained. The VS experiment is part of the work of the EU FP7 funded GRANATUM project (ICT-2009.5.3) [33].

The assessment that follows is a direct result of the experiences and results obtained from the workflows developed. Emphasis is placed on features that make SWMS's attractive, e.g. user friendliness, use of distributed resources, reusability, provenance, etc.

A. User-friendliness

One of the strong selling points of SWMS technology is the promise to allow and trivialize the implementation of complex scientific experiments by non-expert users. Ideally, users with little background in databases and algorithm implementation should be able to design *in silico* experiments that make use of data with varying formats from distributed resources and analyze it using methods executed on computational resources as required. Currently, this is clearly not the general case. Most modern SWMS have made significant steps in this direction but still remain sophisticated tools that can be intimidating to non-computational users. A solution often employed is for SWMS experts to use current technology to implement customized solutions based on user requirements. This custom solution can hide all unnecessary complex details from the end user while at the same time provide equal functionality.

B. Support Mechanisms

The support mechanisms of open source software are typically online resources, such as wiki pages, and mailing lists administered by expert users. As such it is up to the community of each tool to adequately support new users and guide them through their initial usage of the tool. Generally speaking, communities are quick to assist although it usually takes several iterations of email exchanges to solve the problem. In all SWMS examined, more can be done in the form of tutorials, videos, better documentation of common errors, etc.

C. Error Handling

Current SWMS provide tools for error prevention such as data type checking and file data and type checks. Adequate documentation is another tool for the prevention of errors not only to describe what each component or tool does but also to elaborate its inputs, outputs and, common errors and how to deal with them. This is not the case however as common errors or even bugs remain buried into the conversations of mailing lists instead of being updated in the documentation wiki pages. Once an error does happen, the error messages should be meaningful. Unfortunately the systems inspected still produce system related errors incomprehensible to the common user. As such valuable time is lost trying to solve a possibly already solved error.

D. Integration of Heterogeneous Resources

SWMS's have great potential in implementing complex *in silico* experiments integrating computational and data resources from varying sources. Currently, this feature is well supported by some systems such as Galaxy. Support of retrieval of data from online data libraries is an important feature of that tool as is visualizing through online browsers. Other desktop SWMS lack support in this very important feature. Expert users may be able to prepare nodes/processes that communicate with e.g. web services to access and use distributed resources and data repositories but this is not a feature the systems were designed to address or emphasize.

E. Inter-operability

The number of open source SWMS is already large with some being domain specific, others domain independent while still others configured for the grid, for remote services calls, etc. It is also obvious that this number will continue to grow. As noted in [4] the aim is not to restrict the number of SWMS but rather to make sure that these systems can interact. The only feasible way to do this is by the use of standards. Some experts have already argued in favor of using a standard such as the successful business workflow language currently in use. From a user perspective, workflows should be platform independent and specific workflow components and tools should be interoperable. This is not feasible without standardization. An interesting example is the Taverna-Galaxy interoperability. The two systems interact by creating an executable "black box" that encompasses the functionality of the Taverna workflow. The black box can then be executed in the Galaxy platform. Perhaps, the most attractive model for succeeding interoperability is the one the internet is based upon. The workflows packaged as services themselves but this still remains an open research direction.

F. Workflow Sharing

Workflow sharing is one of the key advantages of SWMS. The primary example of well thought workflow sharing can be

found in myExperiment, an online collaboration environment, designed specifically for the sharing of workflows prepared using the Taverna SWMS. Gradually, myExperiment usage is spreading to other open source workflow systems; for example KNIME workflows are also shared through this platform. In the case of Galaxy users can share workflows both within the workbench through the sharing option and by exporting their workflow directly into the myExperiment environment.

G. Provenance Capture

Provenance information ensures the reproducibility of the experiments and as such it plays an important role during the design and execution of a workflow. Provenance is achieved at different levels. The simplest model used keeps track of a workflow execution as a whole. This however does not allow re-execution of a segment of a workflow. Other systems keep track of the data node by node. This makes it easier to make corrections and restart workflow execution for any point of the workflow. Taking into account that scientific workflows are data and processing intensive it is clear that the second model is more suitable.

VI. CONCLUSIONS

Scientific workflows and SWMS have changed the dynamics of many scientific disciplines and have accelerated scientific discoveries. They enable domain scientists to explore, visualize, process, transform, store and model heterogeneous data by the use of functional cross-platform software components utilizing processing power as offered by grid or cloud technology.

In the case of Life Sciences the benefits of SWMS are even greater. The main benefit of the SWMS approach is the transparency it offers for data and resource management. Domain scientists are not interested in the actual form of the data (binary, text, stream or not) nor where the actual processing occurs. They are interested in the results and their presentation form. The second most important benefit is provenance capture. The information gathered by the SWMS during the execution of a workflow enables the reproducibility of the experiment and can also serve for documentation purposes and for future reference. Similarly, with the use of online repositories of scientific workflows documentation is ensured, knowledge flow is promoted and collaboration is encouraged. Finally, current SWMS also provide a friendlier visual working environment supporting easy use.

Online SWMS offer additional benefits. There is no need to set up installations on local machines or remote servers, no downloads, no conflicts and no updates to worry about. All tools are available at any personal computer from anywhere in the world provided that they are connected to the internet. The same applies to data. A scientist can import and use their data in the system available with the workflow to process them. Moreover the data and work are secure and can be backed up and protected depending on user preferences and specific system specifications. Importantly, all data and work can be shared with other collaborators in real time. Some online SWMS even offer more advanced features such as transparent access to HPC, to grid services or the cloud, thus, offering speed and efficiency for scientific processes that are computationally expensive and/or data intensive.

Taken together, the above features can support and accelerate scientific work and discovery. As a result SWMS are gaining ground and are rapidly accepted and used in the daily work routine of numerous research fields. We expect that the next step of the SWMS development will be characterized by wider acceptance and a turn to the online model. Further integration of tools into SWMS systems will surely continue in the immediate future conquering even more scientific domains.

ACKNOWLEDGMENT

The work described has been partially supported by the EU-funded project GRANATUM under FP7-(ICT-2009.5.3) [31].

REFERENCES

[1] "LESC - London e-Science Centre" [Online] Available: http://www.lesc.ic.ac.uk/admin/escience.html, [Accessed : May 01, 2012].

[2] A. Barker and J. I. Van Hemert, "Scientific Workflow: A Survey and Research Directions," PPAM, vol. 4967, no. 5, pp. 746–753, 2007.

[3] T. McPhillips, S. Bowers, D. Zinn, and B. Ludäscher, "Scientific workflow design for mere mortals," Future Generation Computer Systems, vol. 25, no. 5, pp. 541–551, 2009.

[4] B. Ludäscher et al., "Scientific Process Automation and Workflow Management," Development, vol. 10, no. 3, pp. 476–508, 2009.

[5] C. Goble, P. Missier, and D. De Roure, "Scientific workflows," in Current opinion in drug discovery development, vol. 11, no. 3, McGraw Hill, 2008, pp. 527-534.

[6] P. Missier, S. Soiland-Reyes, S. Owen, W. Tan, A. Nenadic, I. Dunlop, A. Williams, T. Oinn, and C. Goble, "Taverna, reloaded," in SSDBM 2010, Heidelberg, Germany, 2010.

[7] D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M. Pocock, P. Li, and T. Oinn, "Taverna: a tool for building and running workflows of services.", Nucleic Acids Research, vol. 34, iss. Web Server issue, pp. 729-732, 2006.

[8] T. Oinn, M. Greenwood, M. Addis, N. Alpdemir, J. Ferris, K. Glover, C. Goble, A. Goderis, D. Hull, D. Marvin, P. Li, P. Lord, M. Pocock, M. Senger, R. Stevens, A. Wipat, and C. Wroe, "Taverna: lessons in creating a workflow environment for the life sciences," Concurrency and Computation: Practice and Experience, vol. 18, iss. 10, pp. 1067-1100, 2006.

[9] M. R. Berthold et al., "KNIME: The Konstanz Information Miner," in Data Analysis Machine Learning and Applications, C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker, Eds. Springer Berlin Heidelberg, 2008, pp. 319-326.

[10] M. R. Berthold et al., "KNIME - the Konstanz information miner: version 2.0 and beyond," SIGKDD Explor Newsl, vol. 11, no. 1, pp. 26-31, 2009.

[11] B. Giardine et al., "Galaxy: A platform for interactive large-scale genome analysis," *Genome Research*, vol. 15, no. 10, pp. 1451-1455, 2005.

[12] D. Blankenberg et al., "Galaxy: a web-based genome analysis tool for experimentalists.," *Current protocols in molecular biology edited by Frederick M Ausubel et al*, vol. Chapter 19, no. January, pp. Unit 19.10.1-21, 2010.

[13] J. Goecks, A. Nekrutenko, and J. Taylor, "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences," Genome Biology, vol. 11, no. 8, p. R86, 2010.

[14] “Pipeline Pilot is Accelrys’ scientific informatics platform.” [Online]. Available: http://accelrys.com/products/pipeline-pilot/. [Accessed: 1-Jul-2012]

[15] A. Rowe, D. Kalaitzopoulos, M. Osmond, M. Ghanem, and Y. Guo, “The discovery net system for high throughput bioinformatics,” Bioinformatics, vol. 19, no. 90001, p. 225i–231, 2003.

[16] Wassermann, B., et al., Sedna: A BPEL-Based Environment for Visual Scientific Workflow Modeling, in Workflows for E-science: Scientific Workflows for Grids, I.J. Taylor, et al., Editors. 2007, Springer-Verlag. p. 428-449.

[17] “Web Services Business Process Execution Language Version 2.0” Internet: http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.pdf, [Accessed: May 01, 2012].

[18] E. Deelman, G. Singh, M.-hui Su, J. Blythe, Y. Gil, and C. Kesselman, “Pegasus□: A framework for mapping complex scientific workflows onto distributed systems,” Jet Propulsion, vol. 13, pp. 219-237, 2005.

[19] “Taverna – open source and domain independent Workflow Management System” [Online]. Available: http://www.taverna.org.uk/. [Accessed: May 01, 2012].

[20] M. Abouelhoda, S. Alaa, and M. Ghanem, “Meta-workflows: pattern-based interoperability between Galaxy and Taverna,” in Proceedings of the 1st International Workshop on Workflow Approaches to New Datacentric Science, 2010, pp. 1-8.

[21] T. Fleuren, J. Götze, and P. Müller, “Workflow Skeletons: Increasing Scalability of Scientific Workflows by Combining Orchestration and Choreography,” 2011 IEEE Ninth European Conference on Web Services, pp. 99-106, 2011.

[22] [1] K. Görlach, M. Sonntag, D. Karastoyanova, F. Leymann, and M. Reiter, “Conventional Simulation Workflow Technology for Scientific Conventional Workflow Technology for Scientific Simulation,” Guide to eScience, pp. 0–31, 2011.

[23] D. D. Roure et al., “myExperiment: Defining the Social Virtual Research Environment,” 2008 IEEE Fourth International Conference on eScience, vol. 0, no. July, pp. 182-189, 2008.

[24] V. Curcin and M. Ghanem, “Scientific workflow systems - can one size fit all?,” 2008 Cairo International Biomedical Engineering Conference, pp. 1-9, 2008

[25] B. Ludäscher, M. Weske, T. McPhillips, and S. Bowers, “Scientific Workflows: Business as Usual?,” BPM, vol. 5701. Springer, pp. 31-47, 2009.

[26] E. Deelman, D. Gannon, M. Shields, and I. Taylor, “Workflows and e-Science: An overview of workflow system features and capabilities,” Future Generation Computer Systems, vol. 25, no. 5, pp. 528-540, May 2009.

[27] I. Taylor, M. Shields, I. Wang, and R. Philp, “Distributed P2P Computing within Triana: A Galaxy Visualization Test Case,” in IPDPS 2003 Conference, 2003, p. 16.1.

[28] I. Altintas, C. Berkley, E. Jaeger, M. Jones, B. Ludascher, and S. Mock, “Kepler: an extensible system for design and execution of scientific workflows,” Proceedings 16th International Conference on Scientific and Statistical Database Management 2004, vol. 59, pp. 423-424, 2004.

[29] R. S. Barga et al., “Trident: Scientific Workflow Workbench for Oceanography,” 2008 IEEE Congress on Services - Part I, pp. 465-466, Jul. 2008.

[30] “KNIME | Konstanz Information Miner.” [Online]. Available: http://www.knime.org/. [Accessed: 1-Jul-2012].

[31] “Galaxy Wiki” [Online]. Available: http://wiki.g2.bx.psu.edu/. [Accessed: 1-Jul-2012].

[32] C. C. Kannas et al., “A Workflow System for Virtual Screening in Cancer Chemoprevention” presented at the IEEE 12th International Conference on BioInformatics and BioEngineering, Larnaka, Cyprus, 2012.

[33] “GRANATUM - Project Vision.” [Online]. Available: http://granatum.org/. [Accessed: 1-Jul-2012].