How Reliable is Your workflow: Monitoring Decay in Scholarly Publications

José Manuel Gómez-Pérez[1], Esteban García-Cuesta[1], Jun Zhao[2], Aleix Garrido[1], José Enrique Ruiz[3]

[1] Intelligent Software Components (iSOCO), Spain.
[2] University of Oxford, UK.
[3] Instituto de Astrofísica de Andalucía, Spain.
jmgomez@isoco.com

Abstract. Scientific workflows play an important role in computational research, as the essential artifacts for communicating the methods used to produce the research findings. We are witnessing a growing number of efforts of treating workflows as first-class artifacts for sharing and exchanging actual scientific knowledge, either as part of scholarly articles or as stand-alone objects. However, workflows are not born to be reliable, which can seriously damage their reusability and trustworthiness as knowledge exchange instruments. Scientific workflows are commonly subject to decaying, which consequently undermines their reliability. In this paper, we propose the hypothesis that reliability of workflows can be notably improved by advocating scientists to preserve a minimal set of information that is essential to assist the interpretations of these workflows and hence improve their reproducibility and reusability. By measuring and monitoring the completeness and stability of this information over time, we are then able to indicate the reliability of scientific workflows, which is critical for establishing trustworthy reuse of these important scientific artifacts.

1 Introduction

Scientific workflows are well-known means to encode scientific knowledge and experimental know-how. By providing explicit and actionable representations of scientific methods, workflows capture such knowledge and support scientific development in a number of critical ways, including the validation of experimental results and the development of new experiments based on the reuse and repurpose of existing workflows. Therefore, scientific workflows are valuable scholarly objects themselves and play an important role for sharing, exchanging, and reusing scientific methods. In fact we are witnessing a growing trend of treating workflows as first-class artifacts, for exchanging and transferring actual scholarly findings, either as part of scholarly articles or as stand-alone objects, as illustrated by popular public workflows repositories like myExperiment [6] and CrowdLabs [15].

Reliability of workflows, i.e. the claimed capability of a workflow, is key to its reuse and as the instrument for knowledge exchange. However, reliability of a workflow can hardly be guaranteed throughout its life time. Scientific workflows are commonly subject to a decayed or reduced ability to be executed or repeated, largely due to the volatility of the external resources that are required for their executions. This is what we call *workflow decay* [20]. Workflow definitions, which record the processes/services used or the data processed, clearly cannot capture all information required to preserve the original capability of the workflows. For example, information about the originator of a workflow is one key piece of information to establish trust on a workflow. A workflow created by a reputable research group or researcher is expected to be more reliable. But this attribution and credit information about workflows may be difficult to address without additional information like provenance metadata about the author.

In order to support these needs for enhancing the reliability of workflows, we propose the adoption of workflow-centric research objects [2] to encapsulate additional information along with workflows, as one single information unit. Such information, structured in the form of annotations following standards like the Annotation Ontology [5], OAI-ORE[1] and PROV-O [2], describes the operations performed by the workflow, provides details on authors, versions or citations, and links to other resources, such as the provenance of the results obtained by executing the workflow, input and output datasets or execution examples. Research objects consequently provide a comprehensive view of the experiment, enable inspection, and support the evaluation of the health of a workflow.

In this paper we propose the hypothesis that reliability of a workflows can be notably improved by preserving a minimal set of essential information along with workflows. This requires a systematics understanding of the causes to workflow decay and hence the set of information to be preserve to prevent or reduce decay. In [20] we produced a classification of causes to workflow decay by systematically analysing a corpus of Taverna workflows selected from the popular public workflow repository, myExperiment.org. Based on our analysis, we identified the minimal set of information to be associated in a workflow to reduce its decay and proposed a minimal information model (Minim) to represent these information as quality requirements that must be satisfied by a research object.

This paper takes a step forward in this direction. Research objects enable scientists to safeguard their workflows against decay by defining and evaluating against a minimal set of requirements that must be satisfied. However, there is a lack of indicators that provide third party scientists with the necessary information to decide whether an existing workflow is reliable or not. Workflows are commonly subject to changes over their life span. On one hand this is due to the nature of knowledge evolution. Workflows are often working scholarly objects that are part of a larger scientific investigation. As scientific understandings develop, workflow designs must be updated accordingly. On the other hand, given the volatile external context that a workflow is built upon, throughout the investigation a workflow may be subject to various changes, to deal with for example, updates of external data formats, data access methods, etc. Our

[1] http://www.openarchives.org/ore/1.0/toc.html
[2] http://www.w3.org/TR/prov-o

method must consider both these internal and external changes when helping the scientists to judge the reliability of a workflow: a workflow that works at the time of inspection cannot be quickly concluded as reliable; while one which does not cannot be simply dismissed as unreliable.

In [20] we introduced the notion of completeness of a research object, i.e., the degree by which a research object contains all the required resources necessary for a purpose (e.g., workflow runnability). In this paper we introduce a new metric, stability, which measures the ability of a workflow to preserve its overall completeness state throughout a given time period. We combine the stability measure with the completeness measure in order to compute the reliability of a workflow. Stability extends the scope of the analysis from a particular point in time to a given time period. Parameters like the impact of the information added or removed from the research object and of the decay suffered throughout its history are taken into account for the computation. In this paper we also present an analytic tool that enables scientists and other stakeholders to visualize these metrics and have a better understanding of the evolution of workflow reliability over time.

The remainder of the paper is structured as follows. Section 2 provides an account of related work relevant for the evaluation of workflow reliability. In section 3 we motivate the need for using completeness and stability measures to establish workflow reliability. We then present an outline of our approach in section 4 and describe our implementation in section 5. In section 6, we illustrate the application of our approach using a case study. Finally, section 7 concludes by summarizing our main contributions and outlining future work.

2 Related Work

Our discussion spans through different areas relevant for scholarly communication dealing with: the modelling of aggregation structure as the basis of new ways of publication and the definition of metrics that assess the information being communicated is conserved free of decay throughout time.

While [14] argued in favor of the use of a small amount of semantics as a necessary step forward in scholar publications, research objects were conceived to extend traditional publication mechanisms [1] and take us beyond the pdf [4] by aggregating essential resources related to experiment results along with publications. This includes not only the data used but also methods applied to produce and analyze those data. The notion of using aggregation to promote reproducibility and accessibility of research has been studied elsewhere, including the Open Archives Initiative Object Reuse and Exchange Specification (OAI-ORE) [18], the Scientific Publication Packages (SPP)[13], and the Scientific Knowledge Objects [8]. Nano-publication [12] is another approach of supporting accessible research by publishing key results of an investigation as concise statements.

Along those lines, an important part of the role of workflow-centric research objects as publication objects is to ensure that the scientific method encoded by a workflow is actually reproducible, therefore providing evidence that the results claimed by the authors actually hold. This has a strong impact in the reuse of workflow-based experiments [9] and is closely related to the goal of myExperiment packs [17], which aggregate elements such as workflows, documents and datasets together, following Web 2.0 and Linked Data principles, in order to support communication and reuse of scientific methods.

In order to enhance the trustworthiness of these ROs we associate them with a list of explicitly defined requirements that they must satisfy and we use this list to evaluate their completeness, i.e. the quality of the ROs with respect to a set of given criteria. This is built upon the idea of a Minimum Information Model (MIM) [7], which provides an encoding of these requirements in OWL[3] and supports reasoning with them. Also related to this is work on information quality in the Web of Data [3] and, more specific to the e-science domain, [16], which focuses on preventing experimental work from being contaminated with poor quality data resulting from inaccurate experiments.

Finally, approaches like [10] aim at validating the execution of specific workflows by checking the provenance of their execution against high level abstractions which act as semantic overlays and allow validating the correct behaviour of the workflow. Complementary work from the field of monitoring and analysis of web-scale service based applications like [11] aims at understanding and analyzing service-oriented applications and eventually detecting and preventing potential misbehaviour.

3 Motivation

To illustrate the need of assessing the reliability of a workflow as a fundamental indicator for reuse, we use an example research object based on a workflow from myExperiment[4] in the Astronomy domain, used to calculate distances, magnitutes and luminosities of galaxies.

In this scenario, Bob has a list of several tens of galaxies that have been observed by members of his group during the last years. He is trying to find a workflow which performs queries on services from the International Virtual Observatory[5] (VO) in order to gather additional complementary physical properties for his galaxies. Related to the tag *extragalactic*, Bob finds a promising workflow in a research object published by Alice. He reads its description and finds some similarities to his problem. He also has a list of galaxies and would like to query several web services to access their physical properties, though not the same as those in Alice's case, and perform similar calculations on them. Bob inspects some of the components of Alice's research object in order to better understand it and to find out what parts he could reuse. Several of the input datasets provided in the research object are interesting, as well as their related information and semantic descriptions.

[3] http://www.w3.org/2004/OWL/
[4] http://www.myexperiment.org/workflows/2560
[5] http://www.ivoa.net

After successfully running the workflow, Bob finally feels confident that Alice's workflow is a perfect candidate for reuse in his own work. However, a deeper anaylsis of its recent history could prove otherwise:

1. The workflow evolution history shows that one of the web services changed the format of the input data when adopting ObsTAP VO[6] standards for multidata querying. As a consecuence the workflow broke, and authors had to replace the format of the input dataset.
2. This dataset was also used in a script for calculating derived properties. The modification of the format of the dataset had consequences in the script, which also had to be updated. Bob thinks this may be very easily prone to errors.
3. Later on, another web service became unavailable during a certain time, which turned out that the service provider (in fact Bob's research institution) forgot to renew the domain and the service was down during two days. The same happened to the input data, since they were hosted in the same institution. Bob would prefer now to use his own input dataset, and not to rely on these ones.
4. This was not the only time the workflow experienced decay due to problems with its web services. Recent replacement of networkign infrastructure (optic fiber and routing hardware) had caused connectivity glitches in the same institution, which is the provider of the web service and input datasets. Bob wonders if he could find another web service to replace this one. He needs his workflow working regularly, since it continuously looks for upgraded data for his statistical study.
5. Finally, very recently a data provider modified the output format of the responses from HTML to VOTable[7] format, in order to be VO compliant and achieve data interoperability. This caused one of the scripts to fail and required the authors to fix it in order to deal with VOTable format instead of proprietary HTML format. Bob thinks this is another potential cause for having scripts behaving differently and not providing good results.

In summary, even though the workflow currently seems to work well, Bob does not feel totally confident about its stability. The analysis shows that trustworthy reuse depends not only on the degree to which the properties of a particular workflow and its corresponding research object are preserved but also on their history. This is especially true for scientists who, like Bob, think a particular workflow can be interesting for them but lack the information about its recent performance. Workflows which can be executed at a particular point in time might decay and become unrunnable in the future if they depend on brittle service or data infrastructure, especially when these belong to third party institutions. Likewise, if they are subject to frequent changes by their author and contributors, the probability that some error is introduced also increases. Therefore, we introduce the concept of workflow stability as a means to consider its recent history an background to evaluate its reliability.

[6] http://www.ivoa.net/Documents/ObsCore
[7] http://www.ivoa.net/Documents/VOTable

4 Approach

We understand *reliability* as a measure of the confidence that a scientist can have on a particular workflow to preserve its capability to execute correctly and produce the expected results. A reliable workflow is expected not only to be free of decay at the moment of being inspected but also in general throughout its life span. Consequently, in order to establish the reliability of a workflow it becomes necessary to assess to what extent it is complete with respect to a number of requirements and how stable it has been with respect to such requirements historically. Therefore, we propose *completeness* (already introduced in [20]) and *stability* as the key dimensions to evaluate workflow reliability. Figure 1 schematically depicts the reliability concept as a three-tiered compound on top of completeness and stability along time.

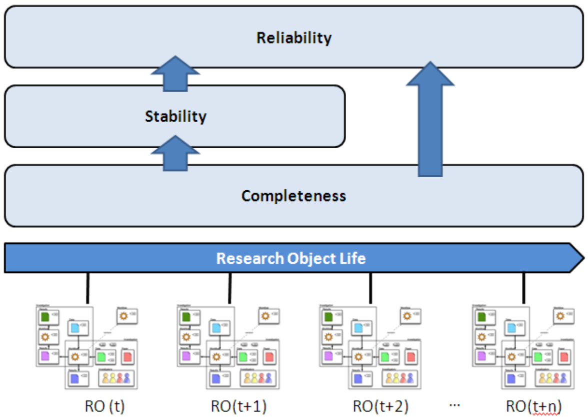

Fig. 1. Layered Components of Reliability Measurement

Following the figure, next sections define each dimension and the relations between them, from completeness to stability and finally reliability.

4.1 Completeness

The completeness dimension evaluates the extent to which a workflow satisfies a number of requirements specified in the form of a checklist following the Minim model[8]. Such requirements can be of two main types: compulsory (must) or recommendable (should). In order to be runnable and reproducible all the *must* requirements associated to a workflow need to be satisfied while *should* requirements propose a more relaxed kind of constraint. An example of the former is

[8] http://purl.org/minim

that all the web services invoked by the workflow be available and accessible (two of the main causes of workflow decay), while the presence of user annotations describing the experiment would illustrate the former.

Since *must* requirements have a strong impact we have defined two thresholds: a) a lower bound β_l which establishes the maximum value that the completeness score can have in case it does not satisfy all *must* requirements, and b) an upper bound β_u which establishes the maximum value that the completeness score can have given that it satisfies all *should* and *must* requirements.

Therefore if at least a *must* requirement fails the completeness score is in the lower band $[0-\beta_l]$ and otherwise in the upper band $[0-\beta_u]$. We define a normalized value of the completeness score as:

$$completeness\_score(RO,t) = f(RO_{(t)}, requirements, type) = \alpha\frac{nSReq(RO_{(t)},must)}{nReq(must)} + (1-\alpha)\frac{nSReq(RO_{(t)},should)}{nReq(should)} \in [0,1],$$

where t is the point in time considered, RO the research object that contains the workflow being evaluated, *requirements* the specific set of requirements defined within the RO for a specific purpose, $type \in \{must, should\}$ the category of the requirement, $\alpha \in [0,1]$ is a control value to weight the different type of requirements, $nSReq$ the number of satisfied requirements, and $nReq$ the total number of requirements for the specified type. This definition of the completeness score gurantees the following properties:

- The maximum value possible if a *must* requirement fails is defined by the lower bound β_l.
- The maximum value possible if all requirements are satisfied is defined by the upper bound $\beta_u = 1$.

4.2 Stability

The stability of a workflow measures the ability of a workflow to preserve its properties through time. The evaluation of this dimension provides the needed information to scientists like Bob the astronomer to know how stable the workflow has been in the past in terms of completeness fluctuation and therefore to gain some insight into how predictable its behavior can be in the near future. We define the stability score as follows:

$$stability\_score(RO,t) = 1 - std(completeness\_score(RO,\Delta t) \in [0.5,1],$$

where completeness_score is the measurement of completeness in time t and Δt is the period of time before t used for evalutation of the standard deviation. The stability score has the following properties:

- It reaches its minimum value when there are severe changes over the resources of a workflow for the period of time Δt, meaning that the completeness score is continuously switching from its minimum value of zero (bad completeness) to its maximum of one (good completeness). This minimum value is therefore associated to unstable workflows.

- It has its maximum value when there are not any changes over a period of time Δt, meaning that the completeness score does not change over that time period. This maximum value is therefore associated to stable workflows.
- Its convergence means that the future behavior of the workflow can be predictable and therefore potentially reusable by interested scientists.

4.3 Reliability

The reliability of a workflow measures its ability for converging towards a scenario free of decay, i.e. complete and stable through time. Therefore, we combine both measures completeness and stability in order to provide some insight into the behavior of the workflow and its expected reliability in te future. We define the reliability score as:
$reliability\_score(RO, t) = completeness\_score(RO, t) * stability\_score(RO, t) \in [0, 1]$,

, where RO is the research object, and t the current time under study. The reliability score has the following properties:

- It has a minimum value of 0 when the completeness score is also minimum.
- It has a maximum value of 1 when the completeness score is maximum and the RO has been stable during the period of time Δt
- A high value of the measure is desirable, meaning that the completeness is high and also that it is stable and hence predictable.

5 Implementation: RO-Monitoring Tool

In this section we describe our developed RO-Monitoring tool which implements the criteria of completeness, stability, and reliability as formulated at section 4. The tool is closely based on the Restful checklist service that was previously presented in [20], which evaluates the completeness of a workflow-oriented research object according to quality requirements expressed using the Minim OWL ontology [9]. Our monitoring tool provides a time-based computation of the completeness, stability and reliability scores of an RO, and stores them as metadata of the evaluated RO in order to create the analytics of the changes of these RO properties over time as shown in Figure 2.

In addition of this monitoring service, we also provide a web-based user interface, using Javascript and JQuery libraries. In this user interface, users can easily compare changes of the completeness of an RO between any two time points, and more importantly, access an explanation of these changes. Users can have a quick overview of who has changed what of an RO, which has brought what kind of the impact in terms of reliability.

The RO-Monitoring service makes use of the Research Object Evolution Model (roevo[10]) to provide explanations to any changes occurred over the time

[9] http://purl.org/net/minim#
[10] http://purl.org/wf4ever/roevo

(e.g. a sudden drop in reliability score). Built upon the latest PROV-O standards, the roevo ontology enables the representation of the different stages of the RO life-cycle, their dependencies, changes and versions. The evolution of an RO over a time period can be tracked and accessed by the ro-evo Restful API[11]. Using the RO evolution traces together with the reliability scores, we can offer end users some meaningful explanation for helping them to interpret the reliability variations, like the number of changes, its type, or the author of those changes.

6 Monitoring RO Decay in Practice

This section shows how our RO-Monitoring tool works in practice with the astronomy case study described in section 3. Figure 2 shows the results produced by the RO-Monitoring tool which visualizes the reliability trace of an astronomy workflow based on the completeness scores computed by daily evaluations, and the stability and reliability scores computed on top of them.

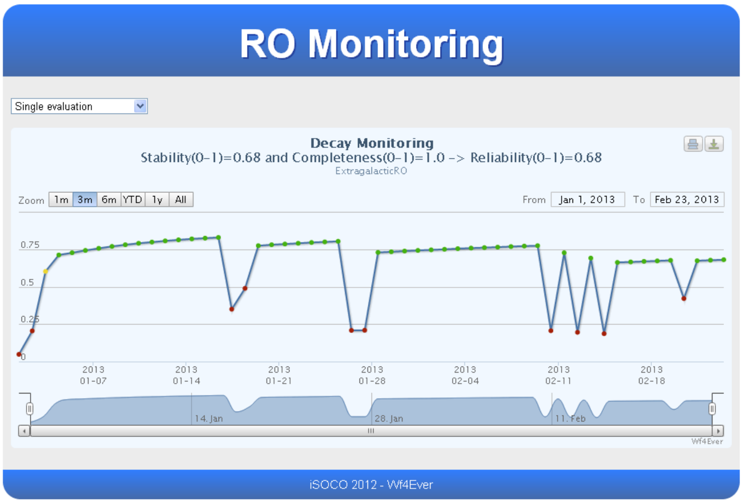

Fig. 2. RO-Monitor web application screenshot for the motivation scenario

Bob wants to reuse a workflow and because a research object contains much richer information for him to reuse the workflow, he starts with such a research object. The first step that he takes is to inspect the RO reliability trace for the RO of his interest. He can see at the beginning of the trace that the RO was initially created some time ago and afterwards its reliability increased due to the addition of some resources. Later on, he observes that there is a first drop

[11] http://sandbox.wf4ever-project.org/decayMonitoring/rest/getAnalytics

on the reliability score, which was caused by a modification of one of the web services that was used by the workflow (i.e. the input format has changed for adopting ObsTAP VO standards). Once the input format is fixed by adopting the standard, the reliability increases; but it still needs more curation by modifying a script that was using the inputs that were changed previously. The second time the reliability drops is due to a period of time where the provider of web services and input data, which turns out to be Bob's institution, has stopped hosting them. When the provider restored the services, the RO reliability recovered and increased along the time until it suffered a successive set of problems related to the services, which were caused again by Bob's institution. This leads to a decrease in the reliability due to these workflow decay problems. The last reliability drop is caused by a script error when a data provider modified its output format from HTML to VOTable.

As we can see, by this reliability trace, Bob can obtain a much more complete picture of the changes of the workflow reliability over a time period, and more interestingly, an explanation behind these changes. This bigger picture as well as the explanations no doubt provide Bob with much more evidence for making decisions about the reliability of the workflow, and hence its reuse.

7 Conclusions and Future Work

Scientists, particularly computational scientists, are demanding new publication paradigms that pay more attention to the methods by which the results reported in publications were obtained. Amongst the various objectives of this movement, it is worthwhile highlighting some of the following, such as the needs for validating the experiment, ensuring that the results are reproducible and therefore trustworthy as the basis of subsequent research, or, more generally speaking, making science more robust, transparent, pragmatic, and useful.

The work presented in this paper falls within such objectives, and in particular it aims at contributing to the conservation and reuse of the published scientific methods. Reliability of these methods plays an important role in reuse. However, indication of reliability cannot be simply drawn based on face value. We show the evidence that, even in the case they were actually runnable and reproducible at the moment of publication, scientific workflows encoding such methods can experience decay due to different causes. When this happens, the reliability of the workflow, i.e. its claimed capability, could have been seriously undermined without careful consideration.

In this paper, we present our approach that is able to provide a more complete picture of changes that may occur to a workflow over a time period, to assist scientists to establish a more truthful indication of its reliability. Our results, though preliminary, show evidence that the minimal set of information that we identified as necessary to be associated within a research object can indeed enable us to effectively assess some specific quality dimensions of a workflow at a time point and to monitor the change of this quality measure over a time period. Evidence is also shown that the completeness, stability and reliability metrics presented herein have the right behaviour to provide scientists with the necessary insight on work developed by third party scientists, helping them decide whether or not to reuse such work for their own experiments and future publications.

Our next steps will focus on the deployment and evaluation of the approach at a sufficiently large scale in specific communities of scientists in the domains of Astronomy and Biology. To this purpose, we are collaborating with scientific publishers like Gigascience interested in the application of our methods and tools in their publication pipeline.

Acknowledgments. The research reported in this paper is supported by the EU Wf4Ever project (270129) funded under EU FP7 (ICT-2009.4.1).

8 References

References

1. S. Bechhofer, I. Buchan, D. De Roure, P. Missier, J. Ainsworth, J. Bhagat, P. Couch, D. Cruickshank, M. Delderfield, I. Dunlop, M. Gamble, D. Michaelides, S. Owen, D. Newman, S. Sufi, and C. Goble. Why linked data is not enough for scientists. Future Generation Computer Systems, 2011.
2. K. Belhajjame, O. Corcho, D. Garijo, J. Zhao, P. Missier, D. Newman, R. Palma, S. Bechhofer, E. García-Cuesta, J.M. Gómez-Pérez, G. Klyne, K. Page, M. Roos, J.E. Ruiz, S. Soiland-Reyes, L. Verdes-Montenegro, D. De Roure, and C.A. Goble. Workflow-centric research objects: First class citizens in scholarly discourse. In Proceeding of SePublica2012, pages 112, 2012.
3. C. Bizer. Quality-Driven Information Filtering in the Context of Web-Based Information Systems. VDM Verlag, 2007.
4. P.E. Bourne, T. Clark, R. Dale, A. De Waard, I. Herman, E. Hovy, D. Shotton, et al. Improving future research communication and escholarship: a summary of findings macquarie university researchonline. 2012. http://force11.org/white paper.
5. P. Ciccarese, M. Ocana, L.J. Garcia Castro, S. Das, and T. Clark. An open annotation ontology for science on web 3.0. J Biomed Semantics, 2(Suppl 2):S4, 2011.
6. D. De Roure, C. Goble, and R. Stevens. The design and realisation of the myexperiment virtual research environment for social sharing of workflows. Future Generation Computer Systems, 25:561567, 2009.
7. David Newman, Sean bechhofer, and David De Roure. myexperiment: An ontology for e-research. In Workshop on Semantic Web Applications in Scientific Discourse in conjunction with the International Semantic Web Conference, 2009.
8. F. Giunchiglia and R. ChenuAbente. Scientific knowledge objects v. 1. Technical report, Technical Report DISI-09-006, University of Trento, 2009.
9. C.A. Goble, D. De Roure, and S. Bechhofer. Accelerating scientists knowledge turns. In Proceedings of The 3rd international IC3K joint conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, 2012.
10. J.M. Gómez-Pérez, O. Corcho. Problem-Solving Methods for Understanding Process executions. Computing in Science and Engineering (CiSE), vol. 10, no. 3, pp. 47-52, May/June, 2008.

11. A. Mos, C. Pedrinaci, G. Alvaro, J.M. Gómez-Pérez, D. Liu, G. Vaudaux-Ruth, S. Quaireau, ServiceWave 2009, in Proceedings of the 2009 International Conference on Service-oriented Computing pp269-282.
12. P. Groth, A. Gibson, and J. Velterop. The anatomy of a nanopublication. Information Services and Use, 30(1):5156, 2010.
13. J. Hunter. Scientific publication packagesa selective approach to the communication and archival of scientific output. International Journal of Digital Curation, 1(1):3352, 2008.
14. P, Lord, S. Cockell, R. Stevens. Three Steps to Heaven: Semantic Publishing in a Real World Workow. In Proceeding of SePublica2012, pages 2334, 2012.
15. P. Mates, E. Santos, J. Freire, and C.T. Silva. Crowdlabs: Social analysis and visualization for the sciences. In SSDBM, pages 555564. Springer, 2011.
16. P. Missier, 2008. Modelling and computing the quality of information in e-science. Ph.D. thesis, School of Computer Science, University of Manchester.
17. David Newman, Sean bechhofer, and David De Roure. myexperiment: An ontology for e-research. In Workshop on Semantic Web Applications in Scientific Discourse in conjunction with the International Semantic Web Conference, 2009.
18. Open archives initiative object reuse and exchange, 2008.
19. Page, K., Palma, R., Houbowicz, P., et. al. (2012). From workflows to Research Objects: an architecture for preserving the semantics of science. In Proceedings of the 2nd International Workshop on Linked Science.
20. J. Zhao, J.M. Gómez-Pérez, K. Belhajjame, G. Klyne, E. García-Cuesta, Garrido A, Hettne K, Roos M, De Roure D, Goble CA. Why Workflows Break - Understanding and Combating Decay in Taverna Workflows. In the proceedings of the IEEE eScience Conference (eScience 2012), IEEE CS, Chicago, USA, 2012.