Workflow-Centric Research Objects: First Class Citizens in Scholarly Discourse

Khalid Belhajjame[1], Oscar Corcho[2], Daniel Garijo[2], Jun Zhao[4], Paolo Missier[9], David Newman[5], Raúl Palma[6], Sean Bechhofer[1], Esteban García Cuesta[3], José Manuel Gómez-Pérez[3], Graham Klyne[4], Kevin Page[4], Marco Roos[7], José Enrique Ruiz[8], Stian Soiland-Reyes[1], Lourdes Verdes-Montenegro[8], David De Roure[4], Carole A. Goble[1]

[1] University of Manchester, UK. [2] Ontology Engineering Group, Universidad Politécnica de Madrid, Spain. [3] iSOCO, Spain. [4] University of Oxford, UK. [5] University of Southampton, UK. [6] Poznan Supercomputing and Networking Center, Poland. [7] Leiden University Medical Centre, Netherlands. [8] Instituto de Astrofísica de Andalucía, CSIC, Spain. [9] Newcastle University, UK.
khalidb@cs.man.ac.uk

Abstract. A workflow-centric research object bundles a workflow, the provenance of the results obtained by its enactment, other digital objects that are relevant for the experiment (papers, datasets, etc.), and annotations that semantically describe all these objects. In this paper, we propose a model to specify workflow-centric research objects, and show how the model can be grounded using semantic technologies and existing vocabularies, in particular the Object Reuse and Exchange (ORE) model and the Annotation Ontology (AO). We describe the life-cycle of a research object, which resembles the life-cycle of a scientific experiment.

1 Introduction

Scientific workflows are used to describe series of structured activities and computations that arise in scientific problem-solving, providing scientists from virtually any discipline with a means to specify and enact their experiments [3]. From a computational perspective, such experiments (workflows) can be defined as directed acyclic graphs where the nodes correspond to analysis operations, which can be supplied locally or by third party web services, and where the edges specify the flow of data between those operations.

Besides being useful to describe and execute computations, workflows also allow encoding of scientific methods and know-how. Hence they are valuable objects from a scholarly point of view, for several reasons: (i) to allow assessment of the reproducibility of results; (ii) to be reused by the same or by a different scientist; (iii) to be repurposed for other goals than those for which it was originally built; (iv) to validate the method that led to a new scientific insight; (v) to serve as *live-tutorials*, exposing how to take advantage of existing data infrastructure, etc. This follows a trend that can be observed in disciplines such

as Biology and Astronomy, with other types of objects, such as databases, increasingly becoming part of the research outcomes of an individual or a group, and hence also being shared, cited, reused, versioned, etc. [11]

However, the use of workflow specifications on their own does not guarantee to support reusability, shareability, reproducibility, or better understanding of scientific methods. Workflow environment tools evolve across the years, or they may even disappear. The services and tools used by the workflow may change or evolve too. Finally, the data used by the workflow may be updated or no longer available. To overcome these issues, additional information may be needed. This includes annotations to describe the operations performed by the workflow; annotations to provide details like authors, versions, citations, etc.; links to other resources, such as the provenance of the results obtained by executing the workflow, datasets used as input, etc.. Such additional annotations enable a comprehensive view of the experiment, and encourage inspection of the different elements of that experiment, providing the scientist with a picture of the strengths and weaknesses of the digital experiment in relation to decay, adaptability, stability, etc.

These richly annotation objects are what we call workflow-centric research objects. The notion of Research Object has been introduced in previous work [20, 19, 1] – here we focus on Research Objects that encapsulate scientific workflows (hence workflow-centric). In particular, we build on earlier work on myExperiment *packs*, which are bundles that contain elements such as workflows, documents and presentations [15]. Other related work is presented in Section 2. In this paper we extend that work making the following contributions: we present a model for specifying workflow-centric research objects (Section 3), and show how it is grounded using semantic technologies; and we characterise and define their lifecycle, illustrating how they evolve over time to be augmented with provenance of the workflow results and semantic annotations (Section 4).

2 Related Work

In certain disciplines (e.g., life sciences), scientific communication channels like journals encourage or mandate authors of submitted papers to include information about the methods used to reach the conclusions claimed in the paper. This has the aim of promoting reproducibility and reuse of the scientific results reported on those papers. For example, most 'wet lab' life science journal papers must contain a 'materials and methods' section that describes the details about the experiments that the authors conducted. These journals typically have strict rules about how to formulate these sections, but from a computational point of view it is weakly structured; hence they are still hard for other scientists to discover and reuse.

The practice of conveying computational methods in a standardised and highly structured way has had less time to evolve in many areas of science. Some journals are also encouraging authors to make available the data and software that have been used and produced, that is, to make data and processes used

part of the published work [8]. For example, Bioinformatics[1] considers software availability as an important prerequisite to the acceptance of the paper. And the NASA ADS (Astrophysics Data System)[2] is linking and referencing papers, references to the journal, data behind the plots used in the papers, catalogues of objects used (as URL references), software used (as URL references to the Astrophysics Source Code Library), instrument used to gather the observed/input data, and the proposal submitted to ask for observation time. These are important steps forward to promote sharing and reuse. However, software and data availability may not be sufficient to check the reproducibility of results, as described in the introduction.

As stated in the introduction, our model is built on earlier work on myExperiment packs [15], which aggregate elements such as workflows, documents and datasets together, following Web 2.0 and Linked Data principles [18, 17]. The myExperiment ontology [14], which forms the basis for our research object model, has been designed such that it can be easily aligned with existing ontologies. For instance, their elements can be assigned annotations comparable to those defined by Open Annotation Collaboration (OAC).

One important aspect of our work is that we make use of abstract workflow templates as a means to annotate workflow templates, facilitating workflow specification (as done by Gil *et al.* [6] and Ludascher *et al.* [9]). Scientists describe a workflow by identifying abstract tasks and specifying scientific analyses using semantic concepts from an underlying domain ontology. The specified abstract workflow is then mapped to a concrete workflow using mappings that specify for each task the underlying service operations that can be used for its implementations.

Our work is complementary to the above proposals in the sense that, in addition to semantic annotations of workflows, we exploit provenance of workflow results to describe workflow templates. In this context, similar proposals are CrowdLab [10], which provides users with the means for publishing data as well as workflows and the provenance of their results to promote the reproducibility of such results, Janus [12] and OPMW [5]. Here we leverage semantic technologies and underline the importance of annotations, which we hope will yield a wide adoption of research objects among scientists. Besides, we allow connecting more elements to the workflow: alternative material, alternative web services, bibliography, the proposal that led to the workflow/experiment, etc.

A clear demand from domains such as bioinformatics and astronomy is the ability to understand a workflow, for which elements outside of the workflow are often needed.

3 A Model for Workflow-Centric Research Objects

Our workflow-centric research object model aims at providing support for the description of the scientific processes described in the previous section in a machine

[1] http://bioinformatics.oxfordjournals.org/
[2] http://labs.adsabs.harvard.edu/

processable format, together with the datasets involved, the results obtained, and their provenance information. The research object will be also accompanied with annotations, which will promote the discover-ability, and therefore the reusability of the processes (workflows), as well as enabling third parties to assess the validity and reproducibility of the results.

Figure 1 illustrates a coarse-grained view of a workflow-centric research object, which aggregates a number of resources, namely:

- a workflow template, which defines the workflow;
- workflow runs obtained by enacting the workflow template
- other artifacts which can be of different kinds, e.g., a paper that describes the research, datasets used in the experiments, etc.;
- annotations describing the aforementioned elements and their relationships.

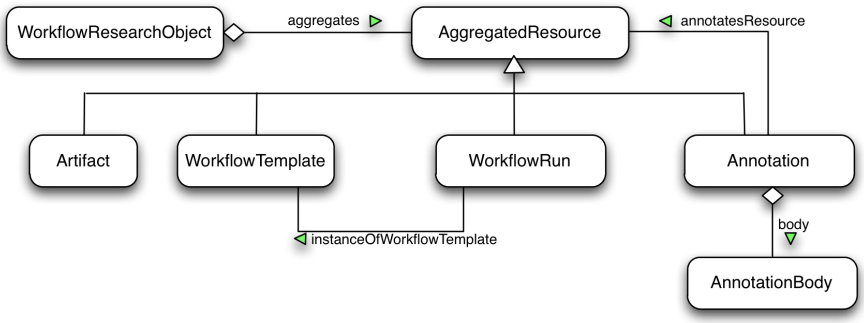

Fig. 1: Workflow-centric research object as an aggregation of resources.

Figure 2 provides a more detailed view of the resources that compose workflow templates and workflow runs. A workflow template is a graph in which the nodes are processes and the edges represent data links that connect the output of a given process to the input of another process, specifying that the artifacts produced by the former are used to feed the latter. A process is used to describe a class of actions that when enacted give rise to process runs. The process specifies the software component (e.g., web service) responsible for undertaking the action. Note that some workflow systems may specify in addition to the data flow, the control flow, which specifies temporal dependencies and conditional flows between processes. We chose to confine the workflow research object model to data-driven workflows, as in Taverna [16], Triana [2], the process run Network Director supplied by Kepler [4], Galaxy[3], Wings [7], etc.

Figure 3-b illustrates an example of a **workflow template** that is composed of two processes. Such a workflow describes an in-silico bioinformatics experiment that is used to identify gene pathways. Specifically, the workflow is composed of two processes: given a protein accession, the *GetKeggGeneId* process is used to

[3] http://galaxy.psu.edu/

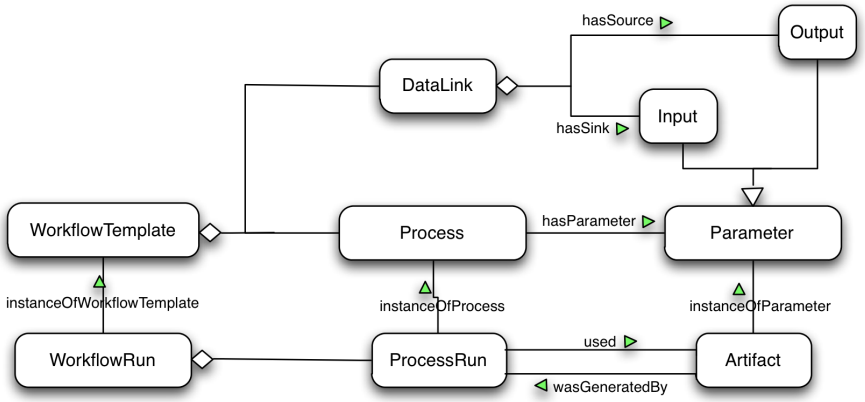

Fig. 2: Resources aggregated within workflow-centric research objects and their relationships.

retrieve the corresponding gene ID. The gene ID retrieved is then used to feed the *GetKeggPathway* process, which returns the corresponding pathways. Note that we also support workflow instances, which are workflow templates with the inputs bound to data values. We also distinguish between standard input parameters and configuration input parameters. Configuration input parameters are used to set the algorithm, the underlying sources used by the processes that compose a workflow template and so on. In addition, the processes that compose a workflow template are not always bound to a software component, rather they can be performed manually in which case they are associated with a human agent.

A workflow template can be instantiated and enacted using a workflow engine, e.g., Taverna. This gives rise to a workflow run that specifies the process runs that were obtained by executing the processes that constitute the workflow template in question. For example, when the action specified by the process is undertaken by a web service, the process run obtained by enacting such a process represents a web service call. A process run may take as input some existing artifacts, specified by the *used* association, and output some new artifacts, specified by the *wasGeneratedBy* association. Artifact is a general concept that represents an immutable piece of state, which may have a physical embodiment in a physical object, or a digital representation in a computer system [13]. In the context of workflow-centric research objects, the focus is on artifacts that are digital representations in a computer system. It is worth mentioning that the notion of process run and artifact that we use are aligned with major provenance models such as the Open Provenance Model (OPM) [13] and PROV-DM[4].

Figure 3-c illustrates an example of a **workflow run** that is obtained by enacting the workflow template together with the provenance of the results produced by the workflow run, which are depicted in Figure 3-b. *GetGeneIdRun*, and *GetGenePathwayRun* are process runs that were obtained

[4] http://www.w3.org/TR/2011/WD-prov-dm-20111018

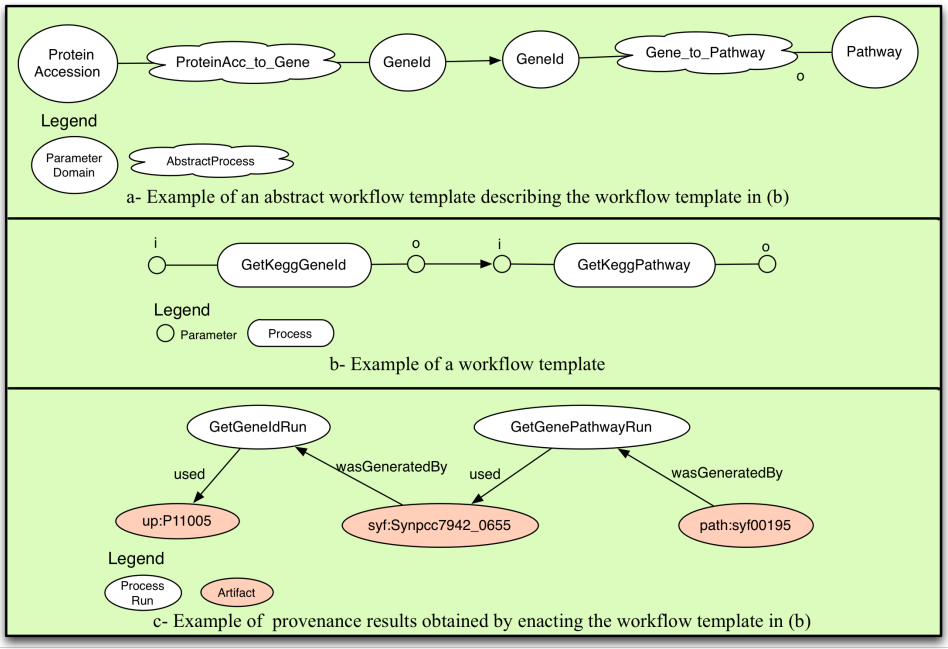

Fig. 3: Example of a workflow template (b), an abstract workflow (a) that semantically describes such workflow template, and provenance of workflow results (c) obtained by enacting the workflow template.

by enacting the *GetGeneId* and *GetGenePathway* processes, respectively. *GetGeneIdRun* took as input the protein accession *up:11005* and generated the gene id *syf:Synpcc7942_0655*, the process run *GetGenePathwayRun* then used *syf:Synpcc7942_0655* to generate the pathway *path:syf00195*.

It is important to highlight that scientists can annotate the elements of a workflow-centric research object (along with the research object itself). They can specify the title of a research object, its purpose, its version, ownership, citations, etc. A more accurate form of annotation can be used to describe the elements of a research object by linking them to concepts from domain ontologies. In particular, this kind of annotation can be used to effectively browse and query workflow templates.

Finally, workflow templates can be annotated in an **abstract workflow template**, which is a graph of abstract processes that are connected by data links. The abstract processes and their input and output parameters are labeled with concepts from underlying domain ontologies, e.g., [21, 22], which specify the tasks performed by the steps and the semantic domains of their parameters, respectively. An abstract worklfow template *awf*, which is used to annotate a given workflow template *wf*, has the same data flow topology as *wf*. The abstract processes that compose *awf* annotate the processes in *wf*, and the parameter domains in *awf* specify the semantic domains of the process parameters in *wf*. As an example, Figure 3-a illustrates an abstract workflow template that semanti-

cally describes the workflow template depicted in Figure 3-b. *ProteinAcc_to_Gene* and *Gene_to_Pathway* are two concepts that specify the tasks of the processes *GetKeggGeneId* and *GetKeggPathway*, respectively, whereas *ProteinAccession*, *GeneId* and *Pathway* are concepts that specify the domain of the input and output parameters of such processes.

3.1 Grounding Workflow-centric Research Objects Using Semantic Technologies

Workflow-centric research objects are encoded using RDF[5], according to a set of ontologies that we have made available[6].

Following myExperiment packs, research objects use the Object Exchange and Reuse (ORE) model[7], to represent aggregation. ORE defines standards for the description and exchange of aggregations of Web resources. Using ORE, a workflow-centric research object is defined as a resource that aggregates other resources, i.e., workflow(s), provenance, other objects and annotations. For example, the RDF turtle snippet illustrated below specifies that a research object identified by `:wro` aggregates a workflow template `:pathway_wf_sp`, a workflow run `:pathway_wf_run`, and an annotation `:wf_annot`.

Example of a research object defined as an ORE aggregation

```
:wro a :WorkflowResearchObject , ore:Aggregation ;
     ore:aggregates :pathway_wf_sp ,
                    :pathway_wf_run ,
                    :wf_annot .
:pathway_wf_sp a :WorkflowTemplate .
:pathway_wf_run a :WorkflowRun .
:wf_annot a ao:Annotation .
```

We also use the Annotation Ontology (AO)[8], which provides a common model for annotating resources. This differs from myExperiment packs, which use a vocabulary that is mapped to Open Annotation Collaboration (OAC)[9][10]. Several types of annotations are supported by the Annotation Ontology, e.g., comments, textual annotations (classic tags) and semantic annotations which relate elements of the research objects to concepts from underlying domain ontologies. As an example, the RDF turtle snippet below shows how the abstract workflow template illustrated in Figure 3-a can be specified using a named graph `:pathway_abs_wf_graph`. It also shows how, using Annotation Ontology, such an abstract workflow template can be used to annotate the workflow template

[5] http://www.w3.org/RDF

[6] http://www.wf4ever-project.org/wiki/display/docs/Research+Object+Vocabulary+Specification

[7] http://www.openarchives.org/ore/1.0/toc.html

[8] http://code.google.com/p/annotation-ontology

[9] www.openannotation.org

[10] Note that work is currently underway to align the two annotation vocabularies: http://www.w3.org/community/openannotation/

`:pathway_wf_sp`, which is depicted in Figure 3-b. Specifically, a resource representing the annotation, `:wf_annot`, is created to link the workflow template which is subject to annotation, `:pathway_wf_sp`, to the named graph specifying the corresponding abstract workflow template, `:pathway_abs_wf_graph`.

Example illustrating how a workflow template can be annotated using AO

```
:wf_annot a ao:Annotation ;
          ao:annotatesResource :pathway_wf_sp ;
          ao:body :pathway_abs_wf_graph.
:pathway_abs_wf_graph {
        :pathway_wf_sp :hasAbsWorkflowTemplate :pathway_abs_wf.
        :pathway_abs_wf a :AbsWorkflowTemplate ;
                :hasAbsProcess :ap1,
                               :ap2 .
                :hasDataLink :dl .
        :ap1 :hasTask :t1 ;
             :hasInput :ap1_in ;
             :hasOutput :ap1_out .
        :t1 a mygrid:ProteinAcc_to_Gene .
        :ap2 :hasTask :t2 ;
             :hasInput :ap2_in ;
             :hasOutput :ap2_out .
        :t2 a mygrid:Gene_to_Pathway .
        :ap1_in :hasDomain :d1 .
        :ap1_out :hasDomain :d2 .
        :ap2_in :hasDomain :d3 .
        :ap2_out :hasDomain :d4 .
        :d1 a mygrid:ProteinAccession .
        :d2 a mygrid:GeneId .
        :d3 a mygrid:GeneId .
        :d4 a mygrid:Pathway .
        :dl :from :ap1_out ;
            :to :ap2_in . }
```

4 The Lifecycle of a Workflow-Centric Research Object

We will now illustrate research object lifecycle through a small example that shows how all the resources contained in a research object are bundled as the scientific experiment progresses. This example lifecycle is summarized graphically in Figure 4.

A research object normally starts its life as an empty **Live Research Object**, with a first design of the experiments to be performed (which determines what workflows and resources will be added, by either retrieving them from an existing platform or creating them from scratch). Then the research object is filled incrementally by aggregating such workflows that are being created, reused or re-purposed, datasets, documents, etc. Any of these components can be changed at any point in time, removed, etc.

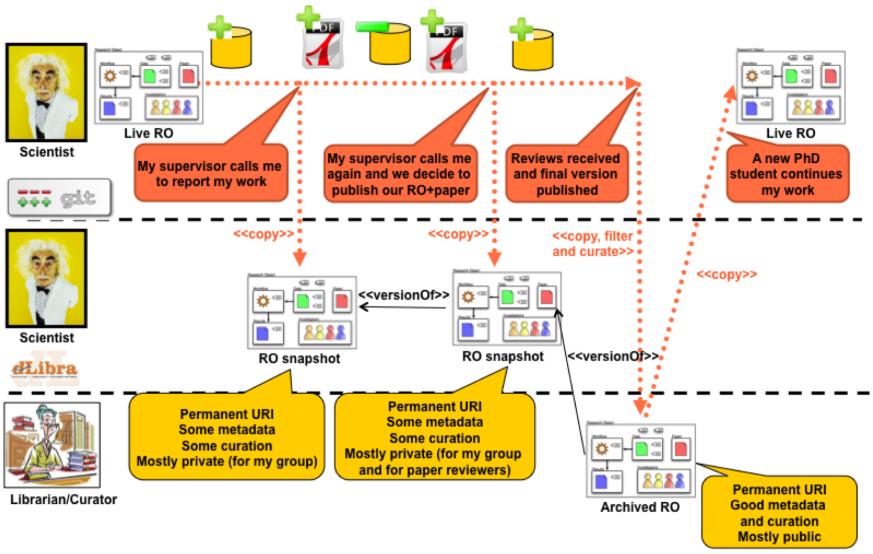

Fig. 4: A sample research object lifecycle.

In our scenario, we observe several points in time when this **Live Research Object** gets copied and kept into a **Research Object snapshot**, which aims to reflect the status of the research object at a given point in time. Such a snapshot may be useful to release the current version of the research outcome of an experiment, submit it to be peer reviewed or to be published (with the appropriate access control mechanisms), share it with supervisors or collaborators, or for acknowledgement and citation purposes.

A snapshot may also contain a paper describing the research object in general and the experiment in particular, depending on the policies of the corresponding scientific communication channel, e.g., workshop, conference or journal. Such snapshots have their own identifiers, and may even be preserved, since it may be useful to be able to track the evolution of the research object over time, so as to allow, for example, retrieval of a previous state of the research object, reporting to funding agencies the evolution of the research conducted, etc.

At some point in time, the research object may get published and archived, in what we know as an **Archived Research Object**, with a permanent identifier. Such a version of our research object may be the result of copying completely our **Live Research Object**, or it may be the result of some filtering or curation process where only some parts of the information available in the aggregation are actually published for others to reuse. As illustrated in Figure 4, a user can use an existing **Archived Research Object** as a starting point to his or her research, e.g., to repurpose it or its parts, in which case a new **Live Research Object** is created based on the existing **Archived Research Object**.

This is only one of the many potential scenarios that could be foreseen for the lifecycle of a workflow-centric research object and we are currently defining different storyboards for their evolution. One important aspect to highlight is

the fact that during its whole lifecycle, the research object is aggregating new objects. The annotation process during the lifecycle of experimentation allows the generation of sufficient metadata about the research objects to support preservation and sharing. Therefore, when a scientists decides to preserve it most of the annotations that will be needed for that preservation process will be already available inside the research object.

5 Conclusions and Further Work

Scientific workflows are used by scientists not only as computational units that encode scientific methods that can be shared among scientists, but also to specify their experiments. In this paper we presented a research object model to capture all the needed information and data including the methods (workflows) and other elements: namely annotations, datasets, provenance of the workflow results, etc.

We showed how this model has been implemented using semantic technologies reusing existing vocabularies, so that scientists are now able to query and publish their experiments according to existing standards. As a result, experiments may be more interoperable, since they are recorded with the same general model to describe them; they can be reused more easily; and decay can be better handled by representing the information of the templates and the traces in an environment/execution independent manner.

The work reported in this paper is preliminary. Our ongoing work includes the design of an architecture for the management of workflow-centric research objects, based on the model presented in this paper, which is being implemented and made available in the Wf4Ever sandbox (http://sandbox.wf4ever-project.eu/). We are also currently validating the model presented in this paper by creating research objects for existing workflows that are stored within the myExperiment repository. In doing so, we are examining issues that have to do with the decay of workflow, mechanisms for querying research objects, and scalability. As well as the technical challenges, we are aware that there are social challenges that need to be overcome to encourage scientists to adopt research object as a unit for publication, discovery and reuse of scientific communications. In this respect, we started collaborating with scientists from the European projects BioVeL (Biodiversity Virtual e-Laboratory)[11] and SCAPE (SCAlable Preservation Environments[12]).

Acknowledgements

The research reported in this paper is supported by the Wf4Ever project (http://www.wf4ever-project.org), Project 270129 funded under EU FP7 Digital Libraries and Digital Preservation (ICT-2009.4.1).

[11] http://www.biovel.eu
[12] http://www.scape-project.eu

References

1. S. Bechhofer, I. Buchan, D. D. Roure, P. Missier, J. Ainsworth, J. Bhagat, P. Couch, D. Cruickshank, and Et Al. Why linked data is not enough for scientists. *Future Generation Computer Systems*, 2011.
2. David Churches and Et Al. Programming scientific and distributed workflow with triana services. *Concurrency and Computation: Practice and Experience*, 18(10):1021–1037, 2006.
3. Ewa Deelman and Et Al. Workflows and e-science: An overview of workflow system features and capabilities. *FGCS*, 25(5):528–540, 2009.
4. Lei Dou and Et Al. Scientific workflow design 2.0: Demonstrating streaming data collections in kepler. In *ICDE*, pages 1296–1299. IEEE Computer Society, 2011.
5. Daniel Garijo and Yolanda Gil. A new approach for publishing workflows: Abstractions, standards, and linked data. In *Proceedings of the Sixth Workshop on Workflows in Support of Large-Scale Science (WORKS'11), held in conjunction with SC 2011*, Seattle, Washington, 2011.
6. Yolanda Gil and Et Al. Mind your metadata: Exploiting semantics for configuration, adaptation, and provenance in scientific workflows. In *International Semantic Web Conference (2)*, pages 65–80. Springer, 2011.
7. Yolanda Gil, Varun Ratnakar, Jihie Kim, Pedro Antonio Gonzalez-Calero, Paul Groth, Joshua Moody, and Ewa Deelman. Wings: Intelligent workflow-based design of computational experiments. *IEEE Intelligent Systems*, 26(1), 2011.
8. Darrel C. Ince, Leslie Hatton, and John Graham-Cumming. The case for open computer programs. *Nature*, 482(7386):485–488, 02 2012.
9. Bertram Ludäscher, Ilkay Altintas, and Amarnath Gupta. Compiling abstract scientific workflows into web service workflows. In *SSDBM*, pages 251–254. IEEE Computer Society, 2003.
10. Phillip Mates, Emanuele Santos, Juliana Freire, and Cláudio T. Silva. Crowdlabs: Social analysis and visualization for the sciences. In *SSDBM*, pages 555–564. Springer, 2011.
11. Jill P. Mesirov. Accessible reproducible research. *Science*, 327(5964):415–416, 2010.
12. Paolo Missier, Satya S Sahoo, Jun Zhao, Carole Goble, and Amit Sheth. Janus: from workflows to semantic provenance and linked open data. *Life Sciences*, 6378(i):129–141, 2010.
13. Luc Moreau and Et Al. The open provenance model core specification (v1.1). *Future Generation Comp. Syst.*, 27(6):743–756, 2011.
14. David Newman. *The Building and Application of a Semantic Platform for an e-Research Society*. PhD thesis, UNIVERSITY OF SOUTHAMPTON, 2011. Submitted on October 2011.
15. 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.
16. Thomas M. Oinn and Et Al. Taverna: lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice and Experience*, 18(10):1067–1100, 2006.
17. Kevin R. Page, David De Roure, and Et Al. Rest and linked data: a match made for domain driven development? In *2nd International Workshop on RESTful Design (WS-REST 2011) held in conjunction with WWW 2011*, 2011.
18. David De Roure and Et Al. The evolution of myexperiment. In *e-Science 2010*. IEEE, 2010.

19. David De Roure, Sean Bechhofer, Carole A. Goble, and David R. Newman. Scientific social objects: The social objects and multidimensional network of the myexperiment website. In *SocialCom/PASSAT*. IEEE, 2011.
20. David De Roure, Khalid Belhajjame, and Et Al. Towards the preservation of scientific workflows. In *Procs. of the 8th International Conference on Preservation of Digital Objects (iPRES 2011)*. ACM, 2011.
21. Vuong Xuan Tran and Hidekazu Tsuji. Owl-t: A task ontology language for automatic service composition. In *ICWS*, pages 1164–1167. IEEE Computer Society, 2007.
22. Chris Wroe, Robert Stevens, Carole A. Goble, Angus Roberts, and R. Mark Greenwood. A suite of daml+oil ontologies to describe bioinformatics web services and data. *Int. J. Cooperative Inf. Syst.*, 12(2):197–224, 2003.