# When History Matters - Assessing Reliability for the Reuse of Scientific Workflows

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

Intelligent Software Components (iSOCO), Spain
{jmgomez,egarcia,agarrido}@isoco.com
Instituto de Astrofísica de Andalucía, Spain
jer@iaa.es
University of Oxford, UK
{jun.zhao,graham.klyne}@zoo.ox.ac.uk

**Abstract.** Scientific workflows play an important role in computational research as essential artifacts for communicating the methods used to produce research findings. We are witnessing a growing number of efforts that treat workflows as first-class artifacts for sharing and exchanging 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 decay, which consequently undermines their reliability over their lifetime. The 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 potential for reproducibility and reusability. In this paper we show how, by measuring and monitoring the completeness and stability of scientific workflows over time we are able to provide scientists with a measure of their reliability, supporting the reuse of trustworthy scientific knowledge.

## 1 Introduction

Workflows have become 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 repurposing of existing workflows. Therefore, scientific workflows 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 findings, either as part of scholarly articles or as stand-alone objects, as illustrated by popular public workflow repositories like myExperiment [5] and CrowdLabs [13].

Workflow reliability, i.e. the capability of a workflow to maintain its properties over time, is key to workflow reuse as the instrument for knowledge exchange.

H. Alani et al. (Eds.): ISWC 2013, Part II, LNCS 8219, pp. 81–97, 2013.

However, workflow reliability can hardly be guaranteed throughout its entire 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* [21].

Workflow definitions, which record the processes/services used or the data processed, clearly cannot capture all the information required to preserve workflows against decay. To this purpose, we propose the adoption of workflow-centric research objects (ROs) [2] to encapsulate additional information along with workflows, as one single information unit. Such information, structured in the form of semantic annotations following standards like the Annotation Ontology [4], OAI-ORE [18] and PROV-O [1], 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. Consequently research objects provide a comprehensive view of the experiment, support the publication of experimental results, enable inspection, and contain the information required for the evaluation of the health of a workflow.

Research objects enable scientists to safeguard their workflows against decay by preserving a minimal set of essential information along with workflows. This requires a thorough understanding of the causes to workflow decay. In [21] we produced a classification of such causes, identified the minimal set of information to be included in a research object, and proposed a minimal information model (Minim) to represent this information as quality requirements that must be satisfied to keep a workflow fit for a purpose (e.g., workflow runnability). We also introduced the notion of completeness of a research object, i.e., the degree by which a research object addresses such requirements.

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 scientific objects that are part of a larger 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 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 this paper we aim at extending the scope of the analysis from a particular point in time to a time period. Parameters like the impact of the information added or removed from the research object and of the decay suffered by the

[1] `http://www.w3.org/TR/prov-o`

workflow throughout its history are taken into account for the computation of its reliability. We formally define the completeness, stability and reliability metrics and propose a lightweight ontological framework, in the context of the Research Object ontologies developed in the Wf4Ever project[2], to support the computation of these metrics. We also present our RO monitoring tool, which implements the approach, enabling scientists to visualize these metrics, analyze the trends, and provide a better understanding of the evolution of workflow reliability over time without requiring a deep knowledge of the underlying knowledge structures.

The remainder of the paper is structured as follows. Section 2 uses a real-life example to motivate the need for combining the completeness and stability metrics to establish a measure of workflow reliability. Section 3 provides an account of relevant related work. Then we present our ontological framework in section 4. Based on such framework, we describe our approach to compute quantitative values of completeness, stability and reliability metrics in section 5. Next, section 6 presents our RO Monitoring tool and provides some implementation details while section 7 illustrates the application of our approach to the motivating example in section 2. Section 8 focuses on the evaluation of our approach with real users in the domain of Astrophysics. Finally, section 9 concludes by summarizing our main contributions and outlining current and future work.

## 2 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[3] in the Astrophysics domain, used to calculate distances, magnitudes and luminosities of galaxies. In this scenario, Bob has a list of several tens of galaxies he has observed during the last years. He is trying to find a workflow that queries the services of the International Virtual Observatory[4] (VO) in order to gather additional 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 and perform similar calculations on them. Bob inspects the research object and, after successfully executing the workflow, feels confident that Alice's workflow is a perfect candidate for reuse in his own work. However, a deeper analysis 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[5] standards for multidata querying. As a consequence the workflow execution broke, and authors had to replace the format of the input dataset.

[2] `http://www.wf4ever-project.org`
[3] `http://www.myexperiment.org/workflows/2560`
[4] `http://www.ivoa.net`
[5] `http://www.ivoa.net/Documents/ObsCore`

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. It turned out that the service provider (in fact Bob's own 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 networking 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 needs his workflow to be run 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[6] 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.

Even though the workflow currently seems to work well, Bob does not feel confident about it. The analysis shows that trustworthy reuse by scientists like Bob 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. Workflows which can be executed at a particular point in time may 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 may increase, too. Therefore, we introduce the stability concept as a means to consider the past history and background of a workflow and evaluate its reliability.

## 3 Related Work

Our discussion spans through different areas dealing with: the modeling of aggregation structures as the basis of scientific information units, especially in the publications domain, and the definition of metrics that assess that the information is conserved free of decay throughout time. While [12] argued in favor of the use of a small amount of semantics as a necessary step forward in scholarly

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

publication, research objects were conceived to extend traditional publication mechanisms [1] 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)[11], and the Scientific Knowledge Objects [7]. Nano-publication [10] is another approach that supports accessible research by publishing key results 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 [8] 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 research objects 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) [6], which provides an OWL encoding of these requirements and supports reasoning with them. Also related to this is work on information quality in the Web of Data [3] and in the e-science domain [14], which focuses on preventing experimental work from being contaminated with poor quality data resulting from inaccurate experiments.

Finally, approaches like [9] 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 their correct behavior. Complementary work from the field of monitoring and analysis of web-scale service-based applications like [15] aims at understanding and analyzing service-oriented applications and detecting and preventing potential misbehavior.

## 4 An Ontological Framework for Reliability Computation

It is not the objective of this paper to provide a complete account of the ontologies developed in the Wf4Ever project to support the modeling of research objects, which are described elsewhere, e.g. in [19]. On the contrary, we will focus on the aspects required to provide the necessary information for establishing a quantitative measure of the reliability, stability, and completeness metrics.

Evaluating the health of the workflow contained in a specific research object requires transforming the additional information encapsulated by the research object into a quantifiable value and providing the scientists with the necessary

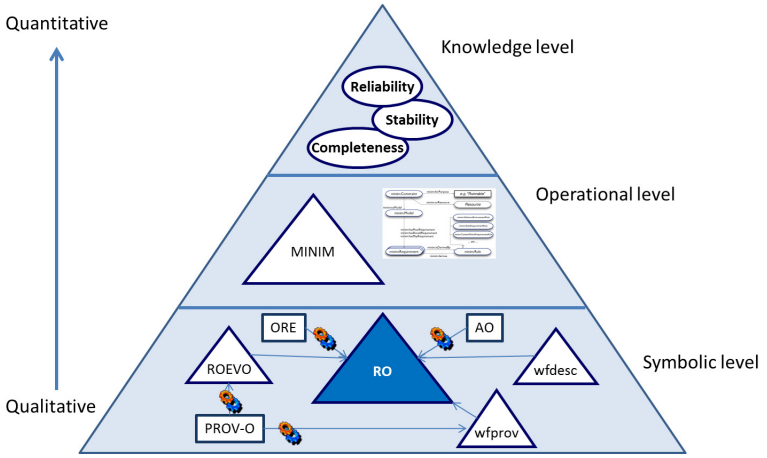


**Fig. 1.** The reliability ontology pyramid

means to interpret such values. We observe a clear separation between the different types of knowledge involved in order to evaluate the reliability of a scientific workflow, as illustrated in Figure 1. Inspired by Newell's knowledge level [16], the figure depicts a pyramid structured in three main layers, where the knowledge about completeness, stability and reliability is obtained through the evaluation of the information contained in the underlying levels.

The bottom layer spans across the main resources included in a research object and can be classified mainly as aggregations of information resources, built on top of the ORE vocabulary, and annotations, following the Annotation Ontology. This layer corresponds to the RO model, described in the RO model specification [19]. This layer is also the placeholder of information related to the workflow included in the research object, in terms of the wfdesc ontology, and of the provenance of its execution, following the wfprov ontology defined as an extension of the PROV-O standard. The Research Object Evolution Ontology (roevo[7]) describes the evolution of research objects over time, providing a record of the changes experienced in the different stages of their lifecycle. Built upon wfprov, the roevo ontology enables the representation of the different stages of the RO life-cycle, their dependencies, changes and versions.

Based on the metadata about the research object, its constituent parts, and annotations, a new layer is included that contains knowledge about the minimum requirements that must be observed by the research object in order to remain fit for a particular goal and about the predicates in charge of evaluating such requirements. This layer, which we call operational in the sense of the methods through which the requirements are evaluated, is modeled as checklists (see [21]) following the Minim OWL ontology[8]. The evaluation of the checklists results

[7] `http://purl.org/wf4ever/roevo`
[8] `http://purl.org/net/minim/minim#`

into a number of boolean values indicating whether the specified requirements are fulfilled or not.

Finally, the top of the pyramid for assessing the reliability of scientific workflows contains quantitative values about reliability, stability, and completeness based on information derived from the outcomes of the checklist evaluation in the previous layer. These metrics are calculated following the algorithms and methods described in section 5 and their values are stored as additional metadata in the research object, providing a compact type of quantitative information about the reliability of specific workflows. Based on these metrics plus the tooling necessary to interpret them (section 7), scientists are enabled to make an informed decision about workflow reuse at the knowledge level, i.e. focusing on their domain expertise and not requiring a deep inspection of the information in the research object.

## 5 Calculating Completeness, Stability and Reliability

We understand *reliability* as a measure of the confidence that a scientist can have in 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 [21]) and *stability* as the key dimensions to evaluate workflow reliability. Figure 2 zooms in the top of the pyramid in Figure 1, schematically depicting the reliability concept as a compound on top of completeness and stability along time.

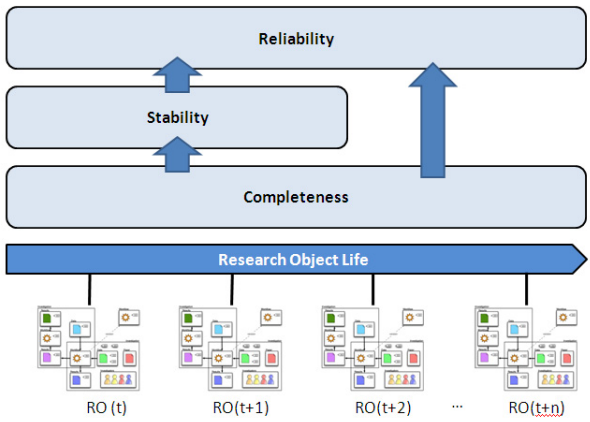


**Fig. 2.** Layered Components of Reliability Measurement

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

### 5.1 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 OWL ontology. 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 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 $\beta_l$ 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 $\beta_u$ which establishes the maximum value that the completeness score can have given that it satisfies all *should* and *must* requirements. Both $\beta_l$ and $\beta_u$ are parameterizable and can be configured on a case by case basis.

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 $[\beta_l - \beta_u]$. Once identified the band, 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 types 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 guarantees the following properties:

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

### 5.2 Stability

The stability of a workflow contributes to measure 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 in order to know

how stable the workflow has been in the past in terms of completeness fluctuation and therefore to gain some insight as to 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 $\Delta t$ is the period of time before $t$ used for evaluation 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 $\Delta t$, 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 $\Delta t$, 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.

### 5.3 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 the 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 $\Delta t$
- A high value of the measure is desirable, meaning that the completeness is high and also that it is stable and hence predictable.

## 6 Implementation: The RO Monitoring Tool

Our monitoring tool provides functionalities for time-based computation of the completeness, stability and reliability scores of an RO, as described in section 5, via a Restful API[9], and stores the results as additional metadata within the RO, as shown in the following sample excerpt of RO metadata in RDF turtle notation. The complete sample RO including this excerpt and the rest of encapsulated metadata, following the RO ontologies [19], and materials can be found in the

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

RO digital library[10] of the Wf4Ever project. The monitoring trace of the RO is available for visualization in the RO Monitoring tool[11].

```
@prefix rdfs:    <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd:     <http://www.w3.org/2001/XMLSchema#> .
@prefix owl:     <http://www.w3.org/2002/07/owl#> .
@prefix rdf:     <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

<http://sandbox.wf4ever-project.org/rodl/ROs/Pack387/>
      <http://purl.org/wf4ever/rovalues#completeness>   1.0 ;
      <http://purl.org/wf4ever/rovalues#reliability>    0.869 ;
      <http://purl.org/wf4ever/rovalues#stability>      0.869 .
```

The resulting information allows producing analytics of the evolution of these metrics over time, as shown in Figure 3. The tool is closely based on the Restful checklist service, previously presented in [21], which evaluates the completeness of a workflow-oriented research object according to quality requirements expressed using the Minim OWL ontology. In addition to the monitoring service, the RO monitoring tool also provides a web-based user interface using JavaScript and jQuery. Through this interface users can inspect the values of these metrics for an RO in time, compare differences between any two time points, and gain access to an explanation of these changes. This allows users to have a quick overview of who has changed what in an RO, and the impact of such actions in terms of reliability. Finally, the RO Monitoring service makes use of the roevo ontology to provide explanations to any changes occurred in a time span, e.g. a sudden drop in the reliability score. Using the RO evolution traces together with the reliability scores, we can offer end users meaningful explanations for helping them to interpret the reliability variations, like the number of changes, its type, or the author of those changes.

## 7 Monitoring Research Object Decay in Practice

Figure 3 shows the reliability trace of the astronomy workflow described in the case study from section 2 produced by our RO monitoring tool. A live demo can be found in the Wf4Ever project sandbox[12]. Our astronomer Bob sees that the RO was initially created some time ago. Soon after new resources were added, with a positive impact in its reliability. He observes that later on there is a first drop on the reliability score, 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). He can inspect the details and compare the status of two different points of the trace in the lower part of the interface. Once the standard is adopted and the input format is fixed, the reliability increases, but further curation is still needed by updating the script using the inputs that were changed previously. The second time reliability drops is framed in a time period

[10] http://sandbox.wf4ever-project.org/rodl/ROs/Pack387

[11] http://sandbox.wf4ever-project.org/decayMonitoring/monitorReliability.html?id=lt

[12] http://sandbox.wf4ever-project.org/decayMonitoring/monitorReliability.html?id=xt

where the infrastructure provider discontinued the hosting of the necessary data and web services. When the provider restored the services, the reliability figures recovered and increased along time until a new set of problems with the same services occurred. The last reliability drop is caused by a script error when a data provider modified its output format from HTML to VOTable.

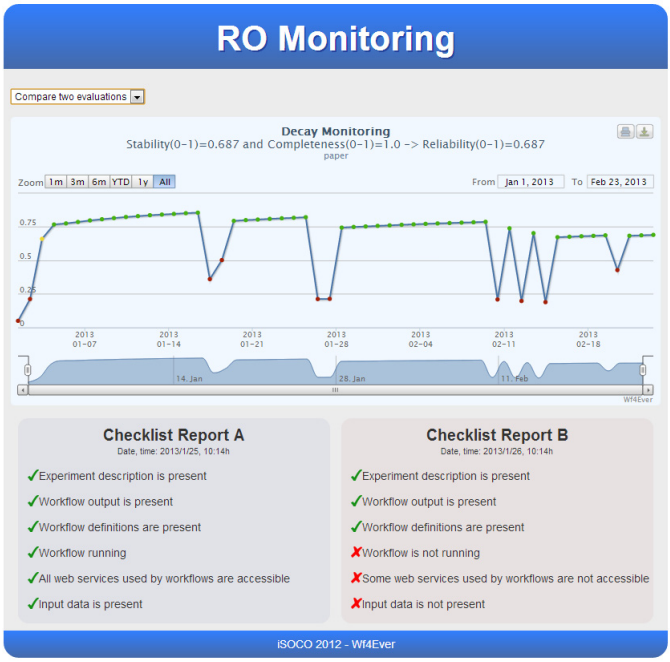


**Fig. 3.** RO-Monitor web application screenshot for the motivation scenario

As shown in the example, our approach provides scientists with indicators of the current reliability of the workflow, based on its general behavior in a particular time period, in order to support decision making for workflow reuse. Of course, it can happen that workflows which have been perfectly stable over their whole lifespan suffer from unexpected decay. In those cases there is not much that can be done apart from following an active monitoring approach for early detection and diagnosis and recording such fluctuation for future references. Under this light, the reliability score reflects the impact of anomalies but prioritizes the importance of the general behavior of the workflow as opposed to isolated events.

## 8 Evaluation

### 8.1 Settings

Collecting the necessary data for evaluating our approach in a real-life setting will require several years after deployment in a production environment like

**Table 1.** Percentage of workflows showing decay per year

| Year | 2007 | 2008 | 2009 | 2010 | 2011 |
|---|---|---|---|---|---|
| Failure % | 91 | 80 | 90 | 50 | 50 |

myExperiment. Though we are taking the necessary steps in this direction, a different, more short-term, approach towards evaluation consists of applying the model of workflow decay that we produced in [21] characterizing workflow decay. In that work we studied the different types of decay in Taverna workflows and obtained real data about the distribution of decay in a four years period. We showed that the most recent workflows are less prone to failures than the older ones, the main explanation being that workflows seem to be no longer maintained after some time since their creation. This makes them less reusable in time, e.g. the amount of workflows created in 2007 suffering from decay was 91% whereas in the case of more recent workflows (2011) it was around 50%.

Following this distribution of workflow decay we have simulated the evolution of 100 workflows during a year, identifying the following three main initial groups of workflows: i) G1 contains the workflow samples which actually run and are well maintained by their creator or any other user with a curator role. G1 workflows are less prone to decay that any other workflow in the other groups; ii) G2 contains those workflows which currently run but are not well maintained by its creator or by a curator. As a consequence G2 workflows can suffer from unexpected decay, especially in the event of changes in external resources necessary for execution; iii) G3 workflows currently do not work properly and there is no guarantee that they will be curated at some point.

In order to model the evolution in time of our workflow population we have considered two different states: an initial state $S1$ at the current time and a final state $S2$ at the end of the sampling period. The distribution of samples considered for each state is obtained from the study in [21]. Table 1 summarizes such figures. The table shows the percentage of decayed workflow for each year, indicating a ratio of decay $rd$ in the end of the fourth year of 39%. We have used this information to establish the initial and final states: the initial state contains 50% workflows that work correctly (according to the data taken from 2011) whereas the final state contains only 9% of the workflows that do so (2007). The distribution of G1, G2 and G3 workflows in the initial and final state of the sample of 100 individuals is (40,93), (20,0) and (40, 7) for each group, respectively.

Given that the initial state converges towards the final state by a constant day probability $P_d$, meaning the likeliness that a workflow changes to another group, we have defined three parameters: $P_d(G1) \propto (1-Stability)$ which establishes the probability that a workflow in G1 is downgraded to G3, $P_d(G2)$ which follows a random distribution for establishing the probability that a workflow in G2 shifts to G1 or G3, and $P_d(G3) \propto Stability$ which establishes the probability that a workflow in G3 is upgraded to G1. For practical reasons we have subsumed G2 into G1 and G3, preserving its individual random behavior. Note that decay

tends to increase as we approach $S2$, hence increasing the population of G3 (Figure 4). The probabilities that a change occurs in a specific day ($P_d$) also follow the analysis in [21]. We have defined $P_d(G1) = 0.49$ and $P_d(G3) = 0.38$, meaning in practice that a workflow will experience three changes of group on average during the year.

Our algorithm implementing this model is shown below in pseudocode. Lines 6 and 10 rank the different workflows of each group proportionally to their stability values (1 - stability for G3); then lines 7 and 11 pick one of them from the 20% first ranked workflows. This ranking method reflects the fact that well maintained workflows will hardly be downgraded from G1 and the opposite for G3 workflows.

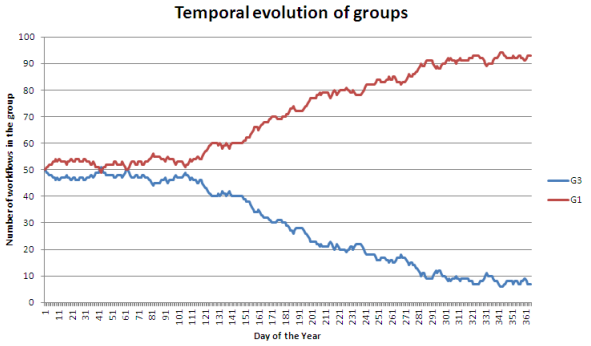


**Fig. 4.** Temporal Evolution of G1 and G3

```
Temporal evolution from S1 to S2
1.     init(G1, G3, State1)
2.     P_d(G1) = 0.49
3.     P_d(G3) = 0.38
4.     for (day = 1 to day == 365)
5.       if (random(1) < P_d(G1) )
6.         rankingG1 = rank(G1, Stability)
7.         toDowngrade = select1From (rankingG1)
8.       end
9.       if (random(1) < P_d(G3) )
10.        rankingG2 = rank(G3, 1-Stability)
11.        toUpgrade = select1From (rankingG2)
12.      end
13.      update(G1, G3, toDowngrade, toUpgrade)
14.    end
```

### 8.2 Evaluation Results

The main objective of this evaluation is to measure the potential benefit for a successful reuse of taking into account a historical perspective on the health of scientific workflows, represented by the reliability score, as opposed to instantaneous quality measures like the completeness value. To this purpose we have run an experiment with nine scientists from the Astrophysics domain [13]. At a given point in time, day 274 of the time simulation, we asked them to look at the completeness values of each of the above mentioned 100 workflows and made them two simple questions: 1. *Would you reuse this workflow for your own experiments today?*, and 2. *Would you use it in three months from now?*. Then, we shuffled

[13] `http://www.iaa.es`

the workflows and asked them to answer the questions again, this time using the RO Monitoring showing the evolution of the reliability of each workflow until day 274. Then we compare both types of results with the actual behavior of each workflow today and in three months.

Two of the users did not pass the control test and were discarded. Thus, we focused on the remaining seven for the evaluation. We have also normalized the results to take into account the subjective point of view of each user. After applying this criteria we made a comparative study between using the completeness and reliability scores, considering the reliability score at the end of the evaluating period, three months ahead, as the ground truth. Our results show that 72% average of the in-the-day reuse decisions (question 1) obtained better results using the reliability score, while this value increased to 76% for question 2. These results are summarized in Table 2. The average distribution for question 1 and 2 for each user was 91%, 85%, 90%, 60%, 75%, 77% and 33%, respectively.

**Table 2.** Reliability vs. Completeness

| | Reuse today | Reuse in 3 months |
|---|---|---|
| Better choice (#times) | 51 | 69 |
| Worse choice (#times) | 19 | 22 |

Furthermore, the reliability score, and its interpretation through the RO monitoring tool, seem to make a better job at managing users' expectations on the convenience of reusing a workflow today or in three months. Based on completeness information alone, 38% workflows would be reused in the day, while incorporating the reliability information constrains this to 32% and even lower (28%) if we ask users to look three months in the future.

Overall we can confirm that the use of the reliability score improves significantly the results obtained using completeness information exclusively. In our experiment we have identified a total of 120 cases where the decision of what workflows should and should not be reused improved using reliability values against 41 negative results. This shows evidence that the use of reliability information, based on the record of workflow health over time, enables scientists to make more informed and better decisions about the reuse of third party scientific workflows, safeguarding their experiments against decay potentially introduced by unstable reused workflows.

## 9 Conclusions and Future Work

Scientists, particularly computational scientists, are paying increasing attention to the methods by which scientific results were obtained. Amongst the advantages that this offers, it is worthwhile highlighting some of the following, such as experimental reproducibility and validation, increased trustworthiness as the basis of subsequent research, and, more generally speaking, making science more robust, transparent, pragmatic, and useful.

The work presented in this paper falls within these lines. In particular we aim at contributing to the conservation and reuse of scientific methods, where reliability plays an important role. However, reliability cannot be drawn simply based on face value. 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 and tool, which are able to provide a more complete picture of the 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 prove 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 specific quality metrics of a workflow at a time point and to monitor the change of this quality measures over a time period. Furthermore, we show how we can obtain compact, quantitative values of those metrics that enable such assessment based on the information stored in the research object encapsulating a scientific workflow.

Our evaluation, conducted by domain experts in the field of Astrophysics, proves that the reliability metric, i.e. considering the combination of workflow completeness and stability in a time period and not just at a single point in time, has a positive impact in the informed reuse of existing workflows by scientists, hence contributing to the development of new workflows based on existing methods. We also provide empiric evidence of how the reliability metric tends to provide a more conservative perspective on the quality of scientific workflows than the completeness metric alone, hence advocating for workflow reuse under safer circumstances. Finally we show that the functions measuring the completeness, stability and reliability metrics presented herein have the right behavior to help scientists decide whether or not to reuse existing work for their own experiments and future work.

We believe our work can have a strong impact in the incremental development of scientific knowledge, especially in those disciplines related to in-silico experimentation, where the reuse of existing work is paramount. New publication paradigms involving semantic publications can benefit from our approach, supporting the development and publication of new scientific advances based on the reuse of reproducible and reliable previous work. To this purpose, we are collaborating with publishers like Gigascience [14] and the American Psychological Association [15] (APA) as well as with scientific digital libraries like NASA's ADS [16] interested in the application of our methods and tools. Other next steps include collecting long-term information about the impact of supporting scientists with information about workflow reliability for a more informed reuse of scientific workflows in the user communities of e-science platforms like

[14] http://www.gigasciencejournal.com
[15] http://www.apa.org
[16] http://adswww.harvard.edu

myExperiment. Through the application of our approach in this scenario we expect a significant increase of the overall quality of the workflows stored in this kind of repositories, where the current amount of unrunnable workflows is currently near to 80% of the total [21] in some cases.

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

## References

1. Bechhofer, S., Buchan, I., De Roure, D., Missier, P., Ainsworth, J., Bhagat, J., Couch, P., Cruickshank, D., Delderfield, M., Dunlop, I., Gamble, M., Michaelides, D., Owen, S., Newman, D., Sufi, S., Goble, C.: Why linked data is not enough for scientists. Future Generation Computer Systems (2011)
2. Belhajjame, K., Corcho, O., Garijo, D., Zhao, J., Missier, P., Newman, D., Palma, R., Bechhofer, S., García-Cuesta, E., Gómez-Pérez, J.M., Klyne, G., Page, K., Roos, M., Ruiz, J.E., Soiland-Reyes, S., Verdes-Montenegro, L., De Roure, D., Goble, C.A.: Workflow-centric research objects: First class citizens in scholarly discourse. In: Proceeding of SePublica 2012, pp. 1–12 (2012)
3. Bizer, C.: Quality-Driven Information Filtering in the Context of Web-Based Information Systems. VDM Verlag (2007)
4. Ciccarese, P., Ocana, M., Garcia Castro, L.J., Das, S., Clark, T.: An open annotation ontology for science on web 3.0. J. Biomed. Semantics 2(suppl. 2), S4 (2011)
5. De Roure, D., Goble, C., Stevens, R.: The design and realisation of the myexperiment virtual research environment for social sharing of workflows. Future Generation Computer Systems 25, 561–567 (2009)
6. Newman, D., Bechhofer, S., De Roure, D.: Myexperiment: An ontology for e-research. In: Workshop on Semantic Web Applications in Scientific Discourse in Conjunction with the International Semantic Web Conference (2009)
7. Giunchiglia, F., ChenuAbente, R.: Scientific knowledge objects v. 1. Technical Report DISI-09-006, University of Trento (2009)
8. Goble, C., De Roure, D., Bechhofer, S.: Accelerating scientists' knowledge turns. In: Fred, A., Dietz, J.L.G., Liu, K., Filipe, J. (eds.) IC3K 2011. CCIS, vol. 348, pp. 3–25. Springer, Heidelberg (2013)
9. Gómez-Pérez, J.M., Corcho, O.: Problem-Solving Methods for Understanding Process executions. Computing in Science and Engineering (CiSE) 10(3), 47–52 (2008)
10. Groth, P., Gibson, A., Velterop, J.: The anatomy of a nanopublication. Information Services and Use 30(1), 51–56 (2010)
11. Hunter, J.: Scientific publication packages – A selective approach to the communication and archival of scientific output. International Journal of Digital Curation 1(1), 33–52 (2008)
12. Lord, P., Cockell, S., Stevens, R.: Three Steps to Heaven: Semantic Publishing in a Real World Workflow. Proceeding of SePublica 2012, 23–34 (2012)
13. Mates, P., Santos, E., Freire, J., Silva, C.T.: Crowdlabs: Social analysis and visualization for the sciences. In: Bayard Cushing, J., French, J., Bowers, S. (eds.) SSDBM 2011. LNCS, vol. 6809, pp. 555–564. Springer, Heidelberg (2011)
14. Missier, P.: Modelling and computing the quality of information in e-science. Ph.D. thesis, School of Computer Science, University of Manchester (2008)

15. Mos, A., Pedrinaci, C., Rey, G.A., Gomez, J.M., Liu, D., Vaudaux-Ruth, G., Quaireau, S.: Multi-level monitoring and analysis of web-scale service based applications. In: Dan, A., Gittler, F., Toumani, F. (eds.) ICSOC/ServiceWave 2009. LNCS, vol. 6275, pp. 269–282. Springer, Heidelberg (2010)
16. Newell, A.: The Knowledge Level. Artificial Intelligence 18(1), 87–127 (1982)
17. Newman, D., Bechhofer, S., De Roure, D.: 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. The Research Object model specification, `http://wf4ever.github.com/ro`
20. Page, K., Palma, R., Houbowicz, P., et al.: From workflows to Research Objects: an architecture for preserving the semantics of science. In: Proceedings of the 2nd International Workshop on Linked Science (2012)
21. Zhao, J., Gómez-Pérez, J.M., Belhajjame, K., Klyne, G., García-Cuesta, E., Garrido, A., Hettne, K., Roos, M., De Roure, D., Goble, C.A.: Why Workflows Break - Understanding and Combating Decay in Taverna Workflows. In: The Proceedings of the IEEE eScience Conference (eScience 2012). IEEE CS, Chicago (2012)