**Stability and Reliability**

The above introduced completeness of a research object provides information of the degree by which a research object contains all the required resources necessary for a purpose (e.g., workflow runnability). Based on this dimension the stability measures the ability of a workflow to preserve its overall completeness state throughout a given period of time. Thereby, 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 its assessment.

We also have defined during the Y3 of the project another quality dimension so called reliability which measures the confidence that a scientist can have in a particular workflow to preserve its capability to be executed correctly and producing the expected results. A reliable workflow is expected not only to be free of decay at the moment of being inspected but also throughout its life. Consequently, in order to establish the reliability of a workflow we have identified 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.

Using these two dimensions we have created an analytic tool (RO monitoring)[[1]](#footnote-1) that enables scientists and other stakeholders to visualize these metrics and have a better understanding of the evolution of workflow reliability over time. This analytic tool also monitors the current stored ROs providing a notification service which alerts whenever a RO suffers of decay.

The formal definition of these dimensions and their scores calculation can be found at D4.2v2 “Design, implementation and deployment of workflow integrity and authenticity maintenance components” and also at [1, 2].

**Component description**

The reliability of a workflow measures the ability of a RO for converging towards a scenario free of decay, i.e. complete and stable through time. 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. For that purpose this component makes used of the different scores for the checklist and stability assessments and then it calculates the reliability of a RO.

The different scores for the completeness dimension are provided by the checklist evaluation API[[2]](#footnote-2), and the stability evaluation is subsumed into the reliability API[[3]](#footnote-3) which provides results for both dimensions simultaneously (see [4] for further details).

**Benchmarking and performance**

The reliability measurement and the implemented visual analytic tool have the main goal of helping end-users (e.g. scientist in the astrophysics and bioinformatics domain) by providing them with a set of indicators which would allow a better judgment whether they want to use one RO or another based on how reliable they are. This criterion has been implemented providing a set of new indicators which have been tested and validated by end users and by usability criteria.

The scenario for validation is the reuse of ROs by scientists. In this scenario a scientist (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 Observatory4 (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 running the workflow, finally 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 VO5 standards for multidata querying. As a consequence 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 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 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 VOTable6 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 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.

Due to collecting the necessary data for evaluating the above introduced scenario using the implemented tools in a real-life setting will require several years after deployment in a production environment (e.g. myExperiment) we have created a scenario which simulates a real one based on empirical data obtained as result of the study done during the Y2 of the project [3]. In that work it was studied the different types of decay in Taverna workflows and obtained real data about the distribution of decay during a period of four years. It was also shown 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%.

We have used this empirical data for characterizing how workflows decay along the time and using a sample of 100 workflows during a year we have identified 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 which has 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 it has been considered two different states: an initial state *S*1 at the current time and a final state *S*2 at the end of the sampling period. The distribution of samples considered for each state was obtained from the study [3]. Table 1 shows the percentage of decayed workflow for each year, indicating a ratio of decay in the end of the fourth year of 41%. We have used this information to establish the initial and final states of the simulation: 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 states of the sample of 100 individuals are (40, 93), (20,0) and (40, 7) for each group, respectively.

Table 1 Percentage of workflows suffering decay per year.

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

Given that the initial state converges towards the final state by a constant day probability Pd, meaning the likeliness that a workflow changes to another group, we have defined three parameters: Pd(G1) ∝ (1- Stability) which establishes the probability that a workflow in G1 (good health) is downgraded to G3 (bad health), Pd(G2) which follows a random distribution for establishing the probability that a workflow in G2 shifts to G1 or G3, and Pd(G3) ∝ 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 (thus having two groups representing a bad and good behavior), although preserving its individual random behavior. Note that decay tends to increase as we approach the final state S2, hence increasing the population of G3 as shown in Fig. 1. The probabilities that a change occurs in a specific day (Pd) also follow the analysis results of [3]. Hereby, we have defined Pd(G1) = 0.49 and Pd(G3) = 0.38, meaning in practice that a workflow will experience three changes of group on average during the year.

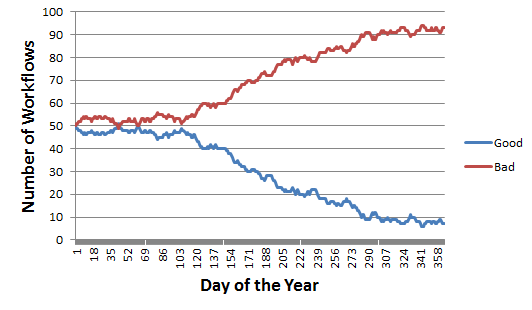


Figure 1 Temporal evolution of the two groups (good and bad behavior)

This simulated scenario was implemented following the above explained model and its pseudocode is shown at Figure 2 where 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 take 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.

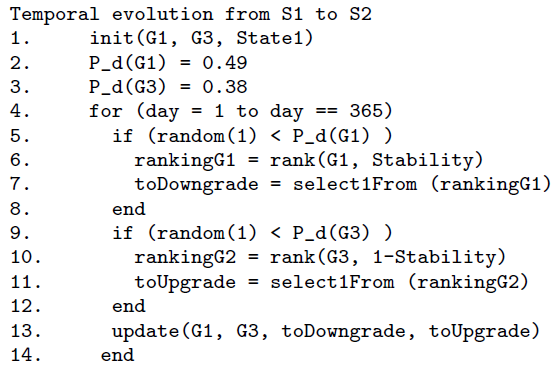


Figure 2 Algorithm for simulating workflows’ behavior evolution

The evaluation of the scenario is done by measuring the potential benefit for a successful reuse 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 run an experiment with nine scientists (as end users) from the Astrophysics domain[[4]](#footnote-4). 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 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 due to the provided answers were outliers. Thus, we focused on the remaining seven for the evaluation. After applying this criterion we made a comparative study between using the completeness and reliability scores, considering the reliability score obtained by the simulation at the end of the evaluating period, three months ahead, as the ground truth. Our results showed 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 improvement distribution for both, question 1 and 2, for each user was 90%, 85%, 90%, 60%, 75%, 77% and 33%, respectively. The complete set of results for the both questions can be seen in tables 3 and 4.

Table 2 Reliability vs. Completeness better choice comparative.

|  |  |  |
| --- | --- | --- |
|  | 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, allow users 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 this future (the complete set of results for all users can be seen at Table 4).

Overall 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 (completeness assessment), 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.

|  |  |  |
| --- | --- | --- |
|  | Better choice (#times) | Worse choice (#times) |
| **User 1** | 10 | 1 |
| **User 2** | 7 | 1 |
| **User 3** | 9 | 1 |
| **User 4** | 6 | 4 |
| **User 5** | 10 | 4 |
| **User 6** | 6 | 2 |
| **User 7** | 3 | 6 |
| **Total** | **51** | **19** |

Table 3 Improvement obtained by using reliability score instead of using completeness score for today question.

|  |  |  |
| --- | --- | --- |
|  | Better choice (#times) | Worse choice (#times) |
| **User 1** | 10 | 1 |
| **User 2** | 10 | 2 |
| **User 3** | 9 | 1 |
| **User 4** | 6 | 4 |
| **User 5** | 19 | 5 |
| **User 6** | 12 | 3 |
| **User 7** | 3 | 6 |
| **Total** | **69** | **22** |

Table 4 Improvement obtained by using reliability score instead of using completeness score for from now in 3 months question.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Completeness (Today) | Reliability (Today) | Reliability (in 3 months) |
| **User 1** | 42 | 31 | 31 |
| **User 2** | 42 | 34 | 30 |
| **User 3** | 42 | 32 | 32 |
| **User 4** | 20 | 20 | 20 |
| **User 5** | 42 | 32 | 20 |
| **User 6** | 42 | 34 | 27 |
| **User 7** | 42 | 43 | 43 |
| **Average** | **37,77 ≈ 38** | **31,61 ≈ 32** | **28,09 ≈ 28** |

Table 5 Number of times users choose to reuse a workflow based on the completeness and reliability tools for questions 1 and 2.

**Performance**

We have tested the reliability service in order to obtain the maximum historical data that can be used without having a penalty in the response to end users which would make the interaction slow. It is known in the literature[[5]](#footnote-5) that the maximum time of response for instantaneous perception is ≈ 0.1-0.2 secs., and between 0.5-1.0 secs. for immediate perception which does not need to communicate any indication to the user to make him aware that the process is ongoing. Considering these values, we have tested how long backwards, in the historical data of a RO, can be represented using the developed RO monitoring tool to provide a instantaneous or immediate perception to end-users (which turns out to be a response between 0 and 1 sec.). The table 5 shows the results obtained regarding the time needed to calculate the reliability scores and response for different number of historical days. If we assume a maximum time of 1 sec. as we just mentioned we could provide results until 2 years and 70 days approximately.

|  |  |
| --- | --- |
| #Days | Response (msecs.) |
| **1** | 38 |
| **10** | 42 |
| **30** | 48 |
| **182** | 64 |
| **365 (1year)** | 105 |
| **730 (2years)** | 912 |
| **1095 (3years)** | 2355 |

Table 6 RO monitoring tool time response for different historical data periods.

**Summary of evaluation for stability and reliability service**

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 implemented stability and reliability service and its RO monitoring visualization provide a new set of indicators which allow scientist and other end users to take better decisions regarding when to reuse a scientific workflow safeguarding their experiments against decay potentially introduced by unstable reused workflows. This implementation has been developed following the Wf4Ever project standards and also trying to accomplish with some characteristics desired by end-users such as usability (e.g. providing user documentation, or reasonable response times), and by developers such as providing releases information, or code documentation.

To test the complete implementation and some of the usability desired characteristics we have run some experiments which simulate real scenarios. These simulations have been based on empirical studies which have provided the values for creating the models of the experiments. The results have shown that using these new tool and the calculated indicators, end-users can take better decision whenever they want to reuse a scientific experiment.

Furthermore we have tested the usability of the interface (RO monitoring tool) regarding the amount of time needed to provide a response. This test has shown that we are currently able to provide a response with an instantaneous or immediate user perception for around 2 years and 2 months of historical data.

**Bibliography**

[1] José Manuél Gómez-Pérez, Esteban García-Cuesta, Jun Zhao, Aleix Garrido, José Enrique Ruiz: How Reliable is Your Workflow: Monitoring Decay in Scholarly Publications. SePublica 2013: 75-86

[2] [José Manuél Gómez-Pérez](http://www.informatik.uni-trier.de/~ley/pers/hd/g/G=oacute=mez=P=eacute=rez:Jos=eacute=_Manu=eacute=l.html), Esteban García-Cuesta, [Aleix Garrido](http://www.informatik.uni-trier.de/~ley/pers/hd/g/Garrido:Aleix.html), [José Enrique Ruiz](http://www.informatik.uni-trier.de/~ley/pers/hd/r/Ruiz:Jos=eacute=_Enrique.html), [Jun Zhao](http://www.informatik.uni-trier.de/~ley/pers/hd/z/Zhao:Jun.html), [Graham Klyne](http://www.informatik.uni-trier.de/~ley/pers/hd/k/Klyne:Graham.html): When History Matters - Assessing Reliability for the Reuse of Scientific Workflows. [International Semantic Web Conference (2) 2013](http://www.informatik.uni-trier.de/~ley/db/conf/semweb/iswc2013-2.html#Gomez-PerezGGRZK13): 81-97

[3] 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.

[4] Esteban García-Cuesta et al. Wf4Ever Technical Report D4.2v2 “Design, implementation and deployment of workflow integrity and authenticity maintenance components”.

1. http://sandbox.wf4ever-project.org/decayMonitoring/visual.html?RO=ro\_uri [↑](#footnote-ref-1)
2. http://www.wf4ever-project.org/wiki/display/docs/RO+checklist+evaluation+API [↑](#footnote-ref-2)
3. http://www.wf4ever-project.org/wiki/display/docs/Reliability+Evaluation+API [↑](#footnote-ref-3)
4. http://www.iaa.es [↑](#footnote-ref-4)
5. <http://www.stevenseow.com/papers/UI%20Timing%20Cheatsheet.pdf> [↑](#footnote-ref-5)