Checklist for Workflow Conservation, v0.1
or
40 rules for Workflow Conservation, v0.1

Please cite as: Garijo D. Corcho O., Belhajjame K. (2014) Checklist for a Workflow Conservation Plan, v0.1. Wf4Ever

Authors:
Daniel Garijo (Facultad de Informática, Universidad Politécnica de Madrid)
Oscar Corcho (Facultad de Informática, Universidad Politécnica de Madrid)
Khalid Belhajjame (LAMSADE, Université Paris-Dauphine)

Sources: This document is largely based on DCC (2013) Checklist for a Data Management Plan, v4.0. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/data-management-plans
and it is also based on the workflow best practice evaluation using checklists described in http://www.wf4ever-project.org/wiki/display/docs/Workflow+best+practice+evaluation+using+checklists

Terminology: There are three different ways to achieve workflow conservation

1) By providing enough metadata about the workflow and about the experiment that the workflow supports, the used and generated datasets, the used software modules or services, etc. That is, a workflow will be sufficiently described, so as to allow other scientists to understand it. In fact, it does not necessarily have to be executable as is (but will have been developed with preservation in mind, what may have influenced the choice of services, datasets, etc.). By describing each step of the workflow, its inputs, outputs, external resources, software being used and intermediate results, we can provide enough information for repairing any possible step that has stopped working. We call this descriptive reproducibility.

2) By making sure that the workflow still executes, using the data input examples that are provided with the original workflow, or datasets selected by a scientist. That is, here we should check whether the workflow is sufficiently specified to at least allow starting its execution, even if the results obtained as a result of the execution are not the same or are even incorrect. We call this workflow execution reproducibility.

3) By making sure that the workflow is properly described, executable and that it produces correct results (which do not necessarily need to be the same as those reported by the published workflow for the same set of results, as long as they are correct). We call this workflow results reproducibility.

The “Result” column is intended for checking whether your workflow has that criterion or not.

N	Criterion	Descriptive rep	Wf exec	Wf rep	Result
1	(**wf**) Specify the purpose/goal of the workflow is documented.	X			
2	(**wf**) Specify the functionality of the workflow (i.e., what the workflow achieves) is documented. For example, a paper describes the workflow.	X			
3	(**wf**) The workflow should have a permanent digital identifier	X	X	X	
4	(**wf**) The workflow should have a diagram or sketch that represents a dataflow of the computational steps is available.	X			
5	(**wf**) Metadata of the workflow describing, amongst other things, the author, creation date, contributors, license, should be provided.	X			
6	(**wf**) When possible, specify alternative services or tools for each workflow step.	X	X		
7	(**wf**) The workflow completes its execution	X	X	X	
8	(**wf**) The workflow delivers the expected results	X		X	
9	(**wf**) The workflow reuses data produced elsewhere	X	X	X	
10	(**wf**) The workflow reuses third party tools/services	X	X	X	
11	(**wf**) All the resources used by the workflow are self contained within the workflow itself (for example, if all the components are scripts or tools included in the specification file itself)		X	X	
12	(**wf**) The workflow has been executed in different software environments.		X	X	
13	(**wf**) The workflow is represented in a format that enables sharing and long term access to its specification and services.	X			

14	(**wf**) The workflow is specified using a standard format	X			
15	(**wf**) The workflow provides the means to read and interpret the specification and results in the future	X			
16	(**wf**) The workflow metadata is defined according to standard specifications and models	X			
17	(**wf**) The workflow assumptions are documented (i.e., requirements to have before executing the workflow, like storing the data somewhere, or reformatting to the appropriate format)	X			
18	(**wf**) The workflow has a license specified.	X			
19	(**wf**/**software**) The restrictions on the workflow for third party data and software are specified.	X	X	X	
20	(**wf**) Owner, creator and creator date of the workflow are provided provided (at least)	X			
21	(**wf**) The intended uses of the workflow are documented.	X			
22	(**wf**) The workflow is stored in a repository or archive.		X	X	
23	(**wf**) The workflow is associated to a service to measure the availability of third party resources and services.		X	X	
24	(**data**)The inputs of the workflow are documented (described)	X			
25	(**data**)The outputs of the workflow are available			X	
26	(**data**)The outputs of the workflow are documented (described)	X			
27	(**data**)The parameters of the workflow are documented (described)	X	X	X	
28	(**data**)The intermediate results of the workflow are available			X	
29	(**data**) The intermediate results of the workflow are documented (described)	X		X	

30	(**data**) Sample data to run the workflow is provided	X	X	X	
31	(**data**) Sample data of a workflow run is available		X	X	
32	(**data**) Input, examples, outputs and intermediate results have a persistent identifier.	X			
33	(**data/software**) All the configuration files used for the tools or external web services are documented (described)	X			
34	(**data/software**) All the configuration files used for the tools or external web services are available		X	X	
35	(**software**)The tool/script/service used for each workflow step is documented in the workflow	X			
36	(**software**) The tool/script/service used for each workflow step is accessible.			X	
37	(**software**) All the scripts / pseudocode used for the workflow are documented (described).	X			
38	(**software**) All the scripts/pseudocode used for the workflow are provided with the workflow itself.			X	
39	(**software**) All workflow steps use open source tools.	X			
40	(**software**) The execution infrastructure of the workflow is documented (i.e., execution environment, software dependencies of the components (like the list of libraries used), physical memory needed to execute the workflow, etc)	X			