Checklist for a Workflow Conservation Plan, v0.1

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

Authors:
Daniel Garijo, Oscar Corcho (Ontology Engineering Group, 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

Checklist	**Guidance and questions to consider**
Administrative Data	
ID	A pertinent ID as determined by the funder and/or institution
Funder	State research funder if relevant
Grant Reference Number	Enter grant reference number if applicable
Project Name	If applying for funding, state the name exactly as in the grant proposal
Project Description	**Questions to consider:** - What is the nature of your research project? - What research questions are you addressing? - For what purpose are the workflows being collected or created? **Guidance:** Briefly summarise the type of study (or studies) to help others understand the purposes for which you are collecting or developing the workflows
PI/Researcher	Name of Principal Investigator(s) or main researcher(s) on the project
PI/Researcher ID	e.g., ORCID
Project Data Contact	Name (if different to above), telephone and email contact details
Date of First Version	Date the first version of the WCP was completed

Date of Last Update	Date the WCP was last changed
Related Policies	**Questions to consider:** - Are there any existing procedures that you will base your approach on? - Does your department/group have data management guidelines? - Does your institution or funder have a workflow conservation plan policy already in place? - Are there any formal standards that you will adopt? **Guidance**: List any other relevant funder, institutional, departmental or group policies on workflow conservation. Some of the information you give in the remainder of the WCP will be determined by the content of other policies. If so, point/link to them here.
Workflow Collection and Creation	
What workflows will you collect and create?	**Questions to consider:** - What type of workflows are you collecting or creating for your experiments? - Are your workflows formally specified in a workflow description language or using a tool? - Are there any existing workflows that you can reuse? **Guidance**: Give a brief description of the workflows that you plan to collect or create, describing the type of language or tools that you will use for development (e.g., Taverna, Wings, Galaxy, etc.). Outline and justify your choice of language or tool, and consider implications of your choice in terms of storage, backup and access.
How will your workflows be collected or created?	**Questions to consider:** - What standards or methodologies will you use? - How will you structure and name the steps and methods used, and the folders and files that you will create for managing the workflow files, the sample input datasets, etc.? - How will you handle versioning? - What quality assurance processes will you adopt? **Guidance**: Outline how the workflows will be developed and which community standards (if any) will be used. Consider how the workflow-related artifacts (workflow files, sample datasets, workflow run provenance logs, etc.) will be organised during the project, mentioning for example naming conventions, version control and folder structures. Explain how the quality of the created workflows will be controlled and documented.
Documentation and metadata	
What documentation and metadata will accompany the workflows?	**Questions to consider:** - What metadata will be included associated to the workflows so as to ensure their (descriptive) reproducibility? - How will that metadata be captured and how will it be created? - What metadata standards will you use and why? **Guidance**:

	Describe the types of metadata and documentation that will accompany the workflows to help other users to understand and reuse it. This may include not only metadata about the workflows, but also about the experiments that the workflows support, the used and generated datasets, the used software modules or services, etc. In fact, the workflows do not necessarily have to be executable as they are, but it is important to describe whether they will be developed with preservation in mind, what may influence 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. A specific checklist with the metadata that is useful to ensure descriptive reproducibility is provided at URL_Checklist.
Preservation	
What is the long-term preservation plan for the workflows?	**Questions to consider:** - Where will the workflows be held (e.g., specialised workflow repositories, general-purpose repositories or archives, etc.)? - What costs if any will your workflow repository or archive charge? - Have you costed in time and effort to prepare the workflow for sharing / preservation? - Are you aiming at descriptive reproducibility, workflow execution reproducibility and/or workflow results reproducibility? **Guidance**: Consider how workflows with long-term value will be preserved and curated beyond the lifetime of the grant. Also outline the plans for preparing and documenting workflows for sharing and archiving. If you do not propose to use an established repository, this document should demonstrate that resources and systems will be in place to enable the workflows to be curated effectively beyond the lifetime of the grant. Describe the level of preservation that you plan to ensure for your workflows: preservation by providing enough metadata to help others understand the experiment and workflow steps, and substitute services if they do not work any more (descriptive reproducibility); preservation of the execution of the workflow using the data input examples (even if the results may not be correct any more); or preservation of the correctness of the results of the workflow (even if they may be different between different executions).
Will you ensure workflow execution reproducibility?	**Questions to consider:** - Will you reuse third party services in your workflows or are all resources used self-contained? - How will you specify the restrictions on third party software and datasets? - Will you test the workflow with different software environments? - Will you provide sample data to run the workflow, and sample runs? **Guidance**: Describe whether you aim at ensuring that the workflow can continue to be executed in the future, and how you will ensure that. To ensure workflow execution reproducibility, the workflow should still work with the data input examples that are provided with the original workflow, or datasets selected by a scientist. Note that the correctness of the results may not be ensured (e.g., an external service starts providing results that are not correct any more). A specific checklist with the metadata that is useful to ensure workflow execution reproducibility is provided at URL_Checklist.

Will you ensure workflow results reproducibility?	**Questions to consider:** - Will you provide the source code of the scripts that are included with the workflows? - Will you provide means to evaluate whether the results obtained by the workflow execution are correct? **Guidance**: Describe whether you aim at ensuring that the workflow not only can continue to be executed in the future, but also still producing correct results, and how you will ensure that. Examples include, besides the previous descriptions, including source code of the scripts that are included with the workflow, so that they can be modified by reusers if needed, providing means to evaluate the correctness of results, etc. A specific checklist with the metadata that is useful to ensure workflow results reproducibility is provided at URL_Checklist.
Storage and backup	
How will the workflows be stored and backed up during the research?	**Questions to consider:** - Do you have to incur charges for workflow storage services, including their intermediate, input or output data? - How will workflows be backed up, and who will be responsible for backup and recovery? - How will workflows be recovered in the event of an incident? **Guidance**: State how often workflows will be backed up and to which locations. Storing workflows, and their associated data, on laptops, computer hard drives or external storage devices alone is very risky. The use of robust, managed storage and backup services provided by university IT teams is preferrable. If you choose to use a third-party services, you should ensure that this does not conflict with any funder, institutional, departmental or group policies, for example in terms of legal jurisdiction in which data is being transferred between workflow or workflow step runs.
How will you manage access and security?	**Questions to consider:** - How you will control access to keep the workflow secure? - How will you ensure the safe transfer of intermediate data between workflow services or steps? **Guidance**: If your workflows, or the data that they are using or generating, are confidential, you should outline any appropriate security measures and note any formal standards that you will comply with.
Workflow Sharing	
How will you share your workflows?	**Questions to consider:** - How will potential users find out about your workflows? - With whom will you share the workflows, and under what conditions? - Will you share the workflows via a repository, on a Website, handle requests directly or use another mechanism? - When will you make the workflows available? - Will you get persistent identifiers for your workflows? **Guidance**: Consider where, how, and to whom workflows with acknowledged long-term value

	should be made available. The methods used to share the workflows will be dependent on a number of factors such as the type, complexity and sensitivity of the workflows that are developed. Consider how people might acknowledge the use of your workflows with other datasets, or the reuse of your workflows in the development of other workflows.
Are any restrictions on workflow sharing required?	**Questions to consider:** - What action will you take to overcome or minimise restrictions? - For how long will you need exclusive use of the workflow and why? - Will a workflow sharing agreement (or equivalent) required? **Guidance**: Outline any expected difficulties in sharing workflows with acknowledged long-term value, along with causes and possible measures to overcome these. Restrictions may be due to confidentiality, authorisation restrictions on the services used, lack of consent agreements or IPR, etc. Consider whether a non-disclosure agreement would give sufficient protection for confidential workflows or their parts.
Ethics and Legal Compliance	
How will you manage any ethical issues?	**Questions to consider:** - Have you gained consent for workflow preservation and sharing? **Guidance**: As with data, ethical issues affect how you store your workflows, who can see/use them and how long they are kept. Managing ethical concerns may include: referral to departmental or institutional ethics committees and formal consent agreements. You should show that you are aware of any issues and have planned accordingly.
How will you manage copyright and Intellectual Property Rights (IPR) issues?	**Questions to consider:** - Who owns the workflows developed? - How will the workflows be licensed for reuse? - Are there any restrictions on the reuse of third-party workflows or third-party data? - Will workflow sharing be postponed/restricted, e.g., to publish or seek patents? **Guidance**: State who will own the copyright and IPR of any workflow that you will collect or develop, along with the license(s) for its use and reuse. For multi-partner projects, IPR ownership may be worth covering in a consortium agreement. Consider any relevant funder, institutional, departmental or group policies on copyright or IPR. Also consider permissions to reuse third-party workflows or data, and any restrictions needed on workflow sharing.
Responsabilities and Resources	
Who will be responsible for workflow conservation?	**Questions to consider:** - Who is responsible for implementing the WCP, and ensuring it is reviewed and revised? - Who will be responsible for each workflow conservation activity? - How will responsibilities be split across partner sites in collaborative research projects?

	- Will workflow ownership and responsibilities for workflow conservation be part of any consortium agreement or contract agreed between partners? **Guidance**: Outline the roles and responsibilities for all activities, e.g., workflow development, metadata production, workflow quality, storage and backup, workflow archiving, etc. Consider who will be responsible for ensuring relevant policies will be respected. Individuals should be named where possible.
What resources will you require to deliver your plan?	**Questions to consider:** - Is additional specialist expertise (or training for existing staff) required? - Do you require hardware or software which is additional or exceptional to existing institutional provision? - Will charges be applied by workflow repositories? **Guidance**: Carefully consider any resources needed to deliver the plan, e.g., software, hardware, technical expertise, etc. Where dedicated resources are needed, these should be outlined and justified.