Flexible Canonical Workflow Frameworks for Science (CWFS)

Philipp Wieder, Peter Wittenburg, Dirk Betz, Claudia Biniossek, Larry Lannom, Christophe Blanchi, George Strawn, Fenghong Liu, Dieter van Uytvanck, Alex Hardisty, Dimitris Koureas, Martin Zünkler, Martin Jaekel, Rainer Stotzka, Per Öster, Jian Hui, Keith Jeffery, Alessandro Spinuso, Daniel Mallmann, Hannes Tiemann, Stan Ahalt, more to come

# Intention

The usage of workflow mechanisms in data science has a long tradition starting with the use of pipe and batch constructions available in some operating systems such as UNIX. While workflow management systems have greatly evolved in computer science, the reality is that work in the data labs is still mainly guided by manual operations and ad-hoc scripts. Only for some exceptional cases, comprehensive and self-documenting workflow software is being used in practice by data scientists. The appearance of frameworks such as Jupyter could change those practices, but we need to go beyond IT-based solutions and design workflows as a part of a larger global data management and processing architecture with the scientists’ needs as the first priority.

**Despite the many excellent developments on IT-based solutions to support workflows, they did not yet change the practices in the data labs. We do not need to develop yet another technical workflow framework, but need to rethink “scientific workflows” and look for approaches that have the potential to change data practices and make them FAIR compliant.**

With this initiative we want to:

* Get an overview about scientifically meaningful and discipline-crossing building blocks to enable scientists to create, run, and reliably re-run workflows.
* Understand how these workflows can make use of Digital Objects[[1]](#footnote-1) at their core and create FAIR compliant data and metadata.
* Get ideas for a re-usable workflow framework that is easy to use for scientists and frees them from an increasing amount of administrative and organisational tasks.
* Understand how these scientifically defined workflows can be implemented with the help of existing technical workflow frameworks to not re-invent the wheel.

The Data Intelligence Journal has decided to devote a special issue to kick-off the CWFS topic, and we would like invite everyone interested to submit contributions. In parallel, we will organise workshops to further intensify discussions about these goals.

# Motivation

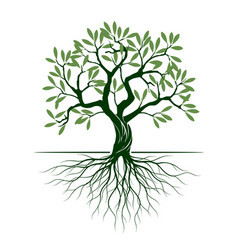
Recently carried out broad surveys indicate that data practices in the labs continue to use traditional approaches such as ad hoc scripts, file systems, database systems, use of non-open file formats, etc. to derive new data or results. This makes it difficult for scientists to create FAIR compliant data and results in the enormous inefficiencies associated with today’s data-driven science, especially when data from different silos/disciplines need to be integrated.

On the other hand, these surveys also indicate that scientists across all disciplines are using a set of typical workflow patterns when:

* generating data by experiments, observations or simulations and after some pre-processing uploading them into trustworthy repositories
* generating new data by combining multiple existing data sets into new collections and then executing some transformation or analysis and storing the result in repositories,
* managing data using typical operations (move, copy, transform, etc.).

It is obvious that, until scientists start using flexible canonical workflow frameworks with integrated FDO compliant operations for all these tasks, the goal of making scientific data easily accessible, understandable and reusable will remain an illusion. However, scientists are fairly conservative and will only accept new methods if they provide clear benefits while not requiring additional work and expense. At the same time scientist are currently confronted with additional legal, ethical, metadata and administrative requirements, to name a few so that workflow frameworks with suitable pluggable components could take over many of these tasks and reduce workload on the scientist[[2]](#footnote-2).

When the RDA initiative was first started it was assumed that only data management would be subject to widely standardised and repeatable operations. However, we now need to realise that the creation and analysis of data, in which scientists are continuously inventing new methods, is also increasingly characterised by repeated operations. Those repeated steps, requiring the scientists’ time and attention, are embedded in many administrative and organisational activities such as requesting permission to run experiments that present sensitive ethical issues to humans, or incrementally creating comprehensive provenance information, all of which impose additional work on scientists which could instead be embedded in semi-automatic workflows.



creation

analytics

management

*This simple diagram was used early in RDA to indicate the potential of harmonizing data management and leaving creation and analysis over to the creativity of scientists. Our understanding has evolved to recognise that these areas can also be generally characterised by repeating actions.*

We are convinced that FAIR Digital Objects and associated supporting infrastructure will be the key to making scientific data persistent, accessible, reusable, and understandable into the next century. But it is clear that scientists and researchers are not interested in any of these implementation details. CWFS has as its goal to hide most of the technical details while creating a framework for automating the creation of scientific data that is DO-based and FAIR compliant. The cost in terms of effort and funding will be high, since developing these CWFS is a complex mix of studying scientific processes, integration work and new development efforts, but given the current huge waste of time, energy, and resources in accessing and/or re-creating scientific data this investment is critical for continuing scientific efficiency and efficacy. These efforts will be cross-disciplinary and also focus on international cooperation.

Of course, there will be exceptions in which scientific processes cannot be carried out with the help of such canonical workflow frameworks and in which researchers will need to continue with ad-hoc scripting etc.

# Canonical Workflows for Lifecycle Support

At its simplest, any workflow can be described as an action that transforms a system from state X to state X+1 as indicated in the figure, which is independent whether we are looking from an object-oriented or actor-oriented point of view. The state of a system can itself consist of various elements such as data sets, conditions, data sources etc. An action consists of the set of processes that consistently transform the system in state X into its state X+1. Each of the processes can themselves be recursively described as their own workflow. A workflow is a concatenation of individual steps that allows for the forking and merging of control paths, can include synchronous and asynchronous activities, letting users to interact with the workflow as well as other existing features available in existing workflow frameworks. Each state within the workflow can be interpreted as a complex collection Digital Object which will include all critical attributes describing the state, as well as a set of qualified PIDs pointing to all other Digital Objects related to the current state (source data and metadata, resulting data and metadata, software that has been used, parameter sets, etc.). As defined by RDA, DOs can be nested, i.e. DOs can be collections that include DOs or other collections. In the case of workflows, one would speak about nested workflows. It should be noted that complex collections describing workflow steps will, in general, include data and operations. The transition from one state to another alsm may be controlled by pre-conditions and post-conditions, a.k.a. constraints.

stateX

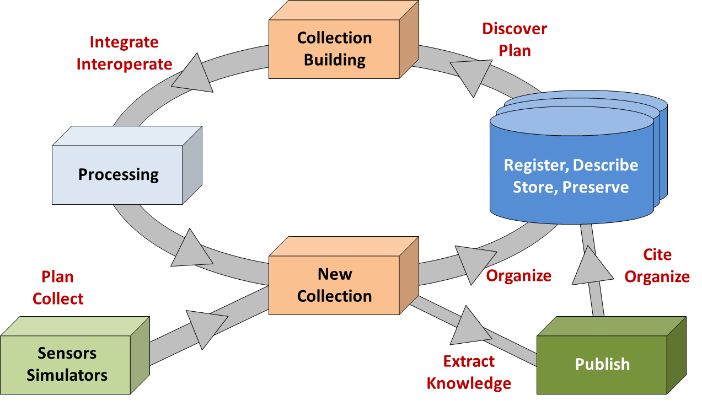
stateX+1

action

CWFS are described at a scientifically meaningful abstraction level which then can be broken down into higher granularity (lower abstraction) levels such as micro-procedures where necessary. To effectively support the researchers the abstraction level needs to employ simple building blocks and terminology familiar to them. Workflows will only find their way into daily practices if the CWFS can be used directly by researchers in as transparent a fashion as possible. There are excellent examples for such scientifically defined workflow types such as Weblicht [] which allows linguists to orchestrate workflows to perform a sequence of linguistic computations without understanding technical details.

Defining such abstract and cross-disciplinary building blocks can be complex and requires studying concrete use cases without being biased by existing technical implementations. From a first overview of many data lab processes we believe that this is will only be possible if building blocks can easily be configured and added, if canonical building blocks are sufficiently abstract so that they can be adapted to specific requirements and if we find ways to describe generic protocols and mechanisms. As indicated, the systematic use of Digital Objects and standardised DO interface protocols will not only simplify the task ahead, but also solve implicitly the request to create FAIR compliant and persistent data and relationships.

# Workflow Frameworks

Workflow frameworks are not new but, in general, the solutions are not made for direct use by researchers not familiar with programming, do not unload researchers from various administrative and organisational tasks, do not inherently produce standardised FAIR Digital Objects without adding a burden on the researcher, do not automate the collection of provenance and curation information, and are not meant to immediately store the DO in trustworthy repositories. The workflows we are aiming at should support the cycle illustrated by the diagram from the RDA Data Fabric IG.

Due to the excellent work by many IT experts early limited workflow tools have been expanded in the meantime to a variety of tools and frameworks that allow not only easy creation of ad-hoc scripting, but also orchestration of complex workflows using high level languages, saving and sharing workflows and their contexts, providing advanced execution frameworks, and many more. The canonical workflow framework defined at a scientific level as intended by this initiative should be neutral with respect to these many technical workflow frameworks, i.e. implementations should make use of these existing frameworks so as to not re-invent the wheel. However, at the beginning of our study we will focus on the scientifically related aspects.

# Concrete Actions

* Organising a video Conference about the intentions November 2020
* Finalising this intention note December 2020
* Forming a core group of experts to steer this work and act

as the Editing Team December 2020

* Inviting experts to contribute to a first special Issue January 2021
* Organising a first (virtual) workshop March 2021
* Deadline for submitting papers May 2021

1. In this note we assume that all DO are FAIR-compliant, thus FAIR DO as defined at the Paris meeting []. [↑](#footnote-ref-1)
2. We should note here that creating typical Data Management Plans is a bureaucratic act while CWFS are productive instruments. [↑](#footnote-ref-2)