Canonical Workflow Framework for Research
CWFR

RDA DFIG Meeting January 2021

CWFR Co-Chairs: Limor Peer, Maggie Hellström, Alex Hardisty, Peter Wittenburg
CWFR Group: about 45 experts from across disciplines

Peter Wittenburg
Max Planck Compute & Data Facility
RDA GEDE
Basic Paradox No1

No doubt for me:
• Automated Workflows will determine data science practices in future.
• Much has been done by IT folks to develop excellent Workflow Technology.

Deep Insights from about 75 RI reports:
• Reality is that only few make use of workflows that go beyond simple scripts.
• Data scientists believe that their work is unique, but one can clearly see recurring patterns in data generation, management & analysis.

So:
Something is wrong/missing!
Basic Paradox No2

No doubt for me:
• Currently data scientists spend ~ 80% of the time with „data wrangling“.
• Increasingly more and excellent tools to help researchers not addressing „wrangling“

Deep Insights from about 75 RI reports:
• Researchers are aware of FAIR and support FAIR but hesitate to change practices.
• Researchers are waiting on new generation of tools that will create FAIRness without bothering them.
• Researchers are not really interested in standards or?

So:
Something is wrong/missing!
Social Aspects – of course

You want to orchestrate/execute a workflow? Then first wait on an expert to get it ready.  
*Sorry – but then I prefer scripting.*

*Researchers don’t like dependence from scarce experts.*

You want to do data science? You must be a genious who will not follow routines.

*Well – is this really true in daily data science work?*
- DS need to prepare their work
- DS need to do many iterations to get evidences
- DS need to adhere to increasing formalisms (DMP, FAIR, law/ethics)

*Use genious minds where it is really necessary.*
*Well where is your flexible tool?*
Solution Spectrum (taken from Alex Hardisty)

Flexible supporting framework

- Automated, structured, repeatable workflows
- Problem solving environments
- Virtual research environments
- Virtual laboratories
- Science gateways

I am genius, do everything different

- Personal scripts, bespoke manual steps
- Jupyter notebooks
- Combo bespoke code and calls out to services

- Our workflow technology is not yet tailored for the many data scientists.
- Some young folks are now getting interested in Jupyter programming
- But is this the way to go?
  - Do you have to become a skilled sw developer to do data science?
  - Do you need to become a car builder to drive a car?

- CWFR is pointing into a different direction

Numerous examples e.g.,

Biodiversity Virtual e-Laboratory, 2011 - 2018

Used Apache Taverna. Many others done similar:
- different disciplines
- other workflow engines
CWFR Analysis: Experiments

- colleagues working on experiments with humans identified 9 recurring steps
- from 9 **canonical steps** only 2 are special
- experiment testing and execution often require special micro-timing, hardware & drivers, etc.
- so embed these Micro-Workflows in a flexible & generic framework

Are the canonical steps all the same across institutes/disciplines? NO
- in physics you don’t need an ethical review (normally) – so skip it
- in medicine the regulations for ethics are stricter than in psycholinguistics – so we need specialised **packages**
CWFR Core

- there will be no one CWFS template that would fit all scenarios
- but we can...
  - identify sufficiently similar coherent and recurring patterns
  - identify recurring canonical actions
  - create libraries of such actions
  - create libraries of specialised packages for these actions
CWFR: FDOs are the glue

WF use cases impression: again 1000 flowers blossoming
- many excellent tools out there integrating some WF steps
- there are so many WF frameworks out there
- no compatibility – no interoperability

CWFR
- create a FDO after each action including all relevant information & references
- iterations within a project can easily be done
- exchange between frameworks can also be done via standardised FDO – thus not bound to specific WF technology
- requires structured approach and typed attributes for machine actionability
- idea of generic glue not so new:
  - UIMA from IBM – now Apache
  - Research Objects
FDO in a Nutshell

- FDOs are machine actionable self-standing 1st class citizens on the Internet
- DOs (all from RDA DFT, DTR, Kernel, DFIG)
  - have bit-sequence encoding content
  - have associated rich metadata
  - have assigned a PID (GUPRI)
- PIDs
  - are persistently resolving to „state“ information based on typed attributes
  - including items such as Type, CTKS, Version, etc. and many references to other FDO such as MD, Rights, BC, etc.
- Types
  - registered in open registries
  - allowing to define operations
- FAIR DO
  - attributes are machine actionable
  - metadata is community defined – not yet machine actionable

FDO support
- abstraction (any content)
- persistent binding
- encapsulation
- a single protocol needed (DOIP)
Benefits of CWFR

- tailored packages for every step can facilitate researchers’ work
- FDOs can be created after each action without bothering the researcher with all details creating the FAIR data domain we are dreaming of
- FDO part: rich metadata can be created cumulatively preserving context & provenance
- standard modules (create metadata, register Handle/DOI, etc.) can be provided
- researcher can go steps back, change a bit and redo steps again
- researchers are not bound to specific technologies
Challenge I

- much has been done on “technical workflow frameworks”
  - CWFR not trying to re-invent the wheel

Library components to be integratable by Jupyter, Galaxy, CWL, etc.

There is much we can build on – we need to check CWFR compliance.
Challenge II

• there is already much software out there facilitating the researchers’ job
• some cover different steps in highly integrated ways
• how to include these often excellent software packages?

• agree on a standardised way to exchange information (FDO!)
• develop adapters
CWFR Activities

• March Working Meeting: check Galaxy framework
• April Working Meeting: check Jupyter framework
• May Working Meeting: check RO Approach – can it be turned to FDOs
• June WM: invite ideas

• Call for Papers for special Issue of Data Intelligence Journal (MIT)
  • deadline 1. March
  • inviting all researchers to contribute to CWFR discussion

• Look for/at OSF CWFR: https://osf.io/2cy86/

• You are welcome to join the discussions
Thanks for the attention.

DO FDO  
GEDE GITHUB: https://github.com/GEDE-RDA-Europe/GEDE
CWFR  
OSF CWFR: https://osf.io/2cy86/