The beauty of workflows and models

Workflows for research. Reproducible research.

Dedicated to Jean-Claude Bradley

A pioneer of open access and open notebook science

Professor Carole Goble
The University of Manchester, UK
The Software Sustainability Institute
carole.goble@manchester.ac.uk
@caroleannegoble

RDMF Meeting, Westminster, 20 June 2014
Virtual Witnessing*

Scientific publications have at least two goals:
(i) to announce a result and
(ii) to convince readers that the result is correct

.....

papers in experimental science should describe the results and provide a clear enough protocol to allow successful repetition and extension

Jill Mesirov
Accessible Reproducible Research
Science 22 Jan 2010: 327(5964): 415-416
DOI: 10.1126/science.1179653

Virtual Witnessing*

Scientific publications have at least two goals:
(i) to announce a result and
(ii) to convince readers that the result is correct

…..

papers in experimental science should describe the results and provide a clear enough protocol to allow successful repetition and extension

Jill Mesirov
Accessible Reproducible Research
Science 22 Jan 2010: 327(5964): 415-416
DOI: 10.1126/science.1179653

“An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment, [the complete data] and the complete set of instructions which generated the figures.”

Ince et al The case for open computer programs, Nature 482, 2012

Morin et al Shining Light into Black Boxes, Science 13 April 2012: 336(6078) 159-160

Protocol

Create a gene list in Excel
Go to NCBI
Retrieve FASTA for each gene
DragonDB Blast each sequence
Copy/paste IDs into a spreadsheet
Run Repeat Headers on each sequence
Copy/paste matched sequence into Excel
Run MacVector cut each seq with EcoRI

Monitor the formation of an acetic anhydride by HMR and CMR in CDC15
1. Make up 1 mL of 1 M solution of propionic and 4-methyl-2-pentanone in CDC15
2. Take HMRs and CMRs of the abedone and aceton. Use 3 sec reaction time and scrape for about 1.5 min for the CMR. This should be proof enough then on the compound at 151 in a million.
3. Continue the two solutions into a 1-dram vial and shake vigorously then transfer to an NMR tube.
4. Take HMR at 5, 10 and 20 minutes after mixing.
5. Take CMR at 25 minutes after mixing.
6. Take HMR at 40 min after mixing.
7. Take CMR at 45 min after mixing.
8. Take HMR at 60 min after mixing.
9. Take CMR at 90 min after mixing.
10. Continue to take NMRs after interval doubling until no more change is observed.
Biodiversity
marine monitoring and health assessment
ecological niche modelling

Enclosed sea problem
(Ready et al., 2010)

Pilumnus hirtellus

Data Intensive Science
Collaborative Science

Sarah Bourlat

http://www.biovel.eu
Data collection

Data discovery

Data assembly, cleaning, and refinement

Ecological Niche Modeling

Statistical analysis

Insights

Scholarly Communication & Reporting

Heterogeneous, autonomous, distributed third party specialist software services

My small data

Third party general services

Different infrastructures

My big data

Heterogeneous, autonomous, distributed public data sets

BioSTIF open modeller

Analytical cycle
Workflows: capture the steps

- assembly & interoperability
- shielding & optimising
- flexible variant reuse
- pipelines & exploration
- repetition & comparison
- record & set-up
- provenance collection
- report & embed

multi-code and multi-resource experiments
in-house and external workflow management systems

http://www.taverna.org.uk
Scientific Workflow Management Systems

Application

Generalist

Scientific Workflow Management Systems

Specialist

Infrastructure

Galaxy

KNIME

LONI Pipeline

Pilot

VT

Nipype

P-GRADE

RAPID MINER

Kepler

Taverna

R

MATLAB

Python

Pegasus
Systems Biology

Modelling Cycle

- Public Data Acquisition
- Model Analysis
- Model Construction
- Public Data Acquisition
- Biological insight
- Hypothesis Generation
- Experiment and Data Generation
- Experiment Analysis
- Public Data Acquisition
- Biological insight

Experimental

Modelling
Aggregated Content Infrastructure

share and interlinking multi-stewarded, mixed, methods, models, data, samples...

- Metadata
- External Databases
- Data
- Articles
- Models

http://www.seek4science.org
http://www.isatools.org
Preservation Planning & Watch

Continuous preservation management

Long term preservation of digital data. Maintaining scans of newspapers, books, records of data; Metadata maintenance; large and automated.

Preservation Policy: Collection level
Control Policy: Low level actions & constraints

SCOUT

Watch

policies

create/reevaluate plans

PLATO

Planning

deploy plan

monitored actions

Taverna workflows

Operations

execute action plan

Repository

RODA

Environment and users

access ingest harvest

Monitored environment and users

Monitored content and events

http://www.scape-project.eu/
Preservation Planning & Watch

Merge a Preservation Action Plan...

... with an Access Workflow

Execution Workflow

Components

Publish

Use

metricDocument1  metricDocument2

SCAPE_Metric_Document_Joiner

combinedMetricDocument

RealizeDOs

SCAPE_Metric_Document_Joiner

TransformRealizedDigitalObject

WriteDOsToRepository

AssessMetrics

MetricInput

CombineMetricDocuments

TransformMetricDocument

BashShell

XMLProcessing

XPath_Service_1  XPath_Service_2

outputFile  measure_output_b  measure_output_a  measure_input_b  measure_input_a

Workflow input ports

Workflow output ports

metricDocument1  metricDocument2

Workflows (and Scripts and Models) are....

...provenance of data
...general technique for describing and enacting a process
...precise, unambiguous, transparent protocols and records.
...often complex, so they need explaining.
...often challenging and expensive to develop.
...know-how and best practice.
...collaborations.

...first class citizens of research
...support the process of research
Workflow publishing

Integrative Frameworks

Journals

galaxyproject.org/
Reproducibility = Hard Work

Open-Paper

DOI: 10.1186/2047-217X-1-18

>11000 accesses

(GIGA)_n

SCIENCE

Linked to DOI

(GIGA)_n

DB

Data sets

Linked to DOI

(GIGA)_n

Galaxy

by CBIIT

Analyses

Open-Data

DOI: 10.5524/100038

78GB CC0 data

Open-Pipelines

Open-Workflows

DOI: 10.5524/100044

Open-Review

8 reviewers tested data in ftp server & named reports published

Open-Code

>5000 downloads

Code in sourceforge under GPLv3:

http://soapdenovo2.sourceforge.net/


8 reviewers tested data in ftp server & named reports published

Enabled code to being picked apart by bloggers in wiki


[Scott Edmunds]
End to end reproducibility

Use our packages to acquire data (both your own and from various data sources), analyze it, add in your narrative, and generate a final publication in any one of widely used formats such as Word, PDF, or LaTeX. Combine our tools with the rich ecosystem of existing R packages.

Repositories
Libraries
Registries
Data Operations in Workflows in the Wild

Analysis of 260 publicly available workflows in Taverna, WINGS, Galaxy and Vistrails

## Research Method Stewardship

### Management, Publishing, Preservation

<table>
<thead>
<tr>
<th>Workflows &amp; Scripts</th>
<th>Different systems Formats</th>
</tr>
</thead>
<tbody>
<tr>
<td>Services &amp; Codes</td>
<td>Web Services Code Libraries Executables</td>
</tr>
<tr>
<td>Models &amp; Algorithms</td>
<td>Mark-up Languages, Mathematical descriptions Standards</td>
</tr>
<tr>
<td>Standard Operating Procedures</td>
<td>Descriptions Standards</td>
</tr>
</tbody>
</table>

- my experiment
- BioVeL
- BiodiversityCatalogue
- BioCatalogue
- BioModels
- MolMeth.org
Curation

Blue collar

Specialist

Incremental

**JIJIT not JIC**

It looks like you're trying to work. Would you like me to bug you instead?

- Annoy me till my eyes bleed
- Go away please

Enclaves

Flirts

Non-intrusive, Non-invasive, Not invisible
Barriers to Data and Code Sharing in Computational Science

Survey of Machine Learning Community, NIPS (Stodden, 2010):

<table>
<thead>
<tr>
<th>Code</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>77%</td>
<td>Time to document and clean up</td>
</tr>
<tr>
<td>52%</td>
<td>Dealing with questions from users</td>
</tr>
<tr>
<td>44%</td>
<td>Not receiving attribution</td>
</tr>
<tr>
<td>40%</td>
<td>Possibility of patents</td>
</tr>
<tr>
<td>34%</td>
<td>Legal considerations</td>
</tr>
<tr>
<td>30%</td>
<td>Timeliness of publication</td>
</tr>
<tr>
<td>30%</td>
<td>Potential for scientific advancement</td>
</tr>
<tr>
<td>20%</td>
<td>Compatibility with existing software</td>
</tr>
</tbody>
</table>

Victoria Stodden, AMP 2011 http://www.stodden.net/AMP2011/
Special Issue Reproducible Research Computing in Science and Engineering July/August 2012, 14(4)

Howison and Herbsleb (2013) "Incentives and Integration In Scientific Software Production" CSCW 2013.
Making practices
Sustainability
Management planning
Deposition
Long term access
Credit
Journals
Licensing
Open source / access

1st Workshop on Maintainable Software Practices in e-Science – e-Science 2012
Stodden, Reproducible Research Standard, Intl J Comm Law & Policy, 13 2009
Making practices
Sustainability
Management planning
Deposition
Long term access
Credit
Journals
Licensing
Open source / access

1st Workshop on Maintainable Software Practices in e-Science – e-Science 2012
Stodden, Reproducible Research Standard, Intl J Comm Law & Policy, 13 2009
The CRAPL: An academic-strength open source license

Academics rarely release code, but I hope a license can encourage them.

Generally, academic software is stapled together on a tight deadline; an expert user has to coerce it into running; and it's not pretty code. Academic code is about "proof of concept." These rough edges make academics reluctant to release their software. But, that doesn't mean they shouldn't.

Most open source licenses (1) require source and modifications to be shared with binaries, and (2) absolve authors of legal liability.

An open source license for academics has additional needs: (1) it should require that source and modifications used to validate scientific claims be released with those claims; and (2) more importantly, it should absolve authors of shame, embarrassment and ridicule for ugly code.

Openness should also hinge on publication: once a paper is accepted, the license should force the release of modifications. During peer review, the reviewers. If the paper is rejected, the modifications should remain secret to protect the authors' right to priority.
Software release paradigm

Some of your data isn’t data
Not a static document paradigm

- Release research
- Methods in motion.
- Versioning
- Forks & merges
- F1000, PeerJ GitHub....
Pivot around method / software / data rather than paper.

Citation semantics: software as was? software as is?

Code as a Research Object
Mozilla Science Lab

Get credit for your code!
Archive your GitHub code repository to figshare and receive a citable DOI.

GitHub
Edit, share and improve your code in a collaborative environment.

Mozilla Science Lab
Tools to get your research on the web.

figshare
Persistent, citable, long-term archiving for your research outputs.

The multi-dimensional paper
methods, reproducibility

what does it mean for content managers and the research workflow?
Replication Gap

Out of 18 microarray papers, results from 10 could not be reproduced

1. Ioannidis et al., 2009. Repeatability of published microarray gene expression analyses. *Nature Genetics* 41: 14
“When I use a word," Humpty Dumpty said in rather a scornful tone, "it means just what I choose it to mean - neither more nor less."

*Lewis Carroll, *Through the Looking-Glass*, and What Alice Found There (1871)
Drummond C Replicability is not Reproducibility: Nor is it Good Science, online
Drummond C Replicability is not Reproducibility: Nor is it Good Science, online
Can I repeat & defend my method?

Can I review / replicate and certify your method?

Can I review / reproduce and compare my results / method with your results / method?

Can I transfer your results into my research and reuse this method?

* Adapted from Mesirov, J. Accessible Reproducible Research Science 327(5964), 415-416 (2010)
Ten Simple Rules for Reproducible Computational Research

1. Computational experiments should be recomputable for reproducibility.
2. Recomputation of recomputable experiments should be automated.
3. It should be easier to make experiments recomputable than not.
4. Tools and repositories can help recomputation become standard.
5. The only way to ensure recomputability is to provide virtual machines.
6. Runtime performance is a secondary issue.

Code
All source code written specifically to process data for a published paper must be available to the reviewers and readers of the paper.

Copyright
The copyright ownership and license of any released source code must be clearly stated.

Citation
Researchers who use or adapt science source code in their research must credit the code’s creators in resulting publications.

Credit
Software contributions must be included in systems of scientific assessment, credit, and recognition.

Curation
Source code must remain available, linked to related materials, for the useful lifetime of the publication.

recomputation.org
sciencecodemanifesto.org
Adapted Freire, 2013

Provenance Tracking, Versioning
Replay, Record, Repair

Provenance
Gather dependencies
Capture steps
Track & keep results

Open
Accessible
Available

Description
Intelligible
Machine-readable

Authoring
Exec. Papers
Link docs to experiment

Sweave
LabTrove

Collage
Utopia

github
ProvStore
PROV-PINGS

Sumatra

Build into the workflows of research....

Matrix Population Model analysis v10

Results:
- cohen_cumulative_distance
- confidence_interval_95pc_of_lambda
- damping_ratio

Elasticity matrix

S 0.037
J 0.001 0.011
V 0.003 0.006 0.279 0.003 0.005
G 0.022 0.002 0.197 0.022
D 0.006 0.02 0.002

Sensitivity matrix 1

https://www.youtube.com/watch?v=QVQwSOX5S08

http://nbviewer.ipython.org/github/myGrid/DataHackLLeiden/blob/alan/Player_example.ipynb
RDataTracker and DDG Explorer

```r
read.data <- function(data.file, 
                      start.date, end.date, variable) { 

  ddg.start("read.data")
  zz <- read.csv(data.file)
  ddg.procedure(pname="read.csv",
                 ins=list("data.file"),
                 outs.snapshot=list("zz"))

  zz$date <- as.Date(zz$date)
  all.data <- subset(zz,
                    zz$date>=start.date & zz$date<=end.date)
  ddg.procedure(pname="subset dates",
                ins=list("zz", "start.date", "end.date"),
                outs.snapshot=list("all.data"))

  raw.data <- all.data[,c("date",variable)]
  names(raw.data)[names(raw.data)==variable] <- "raw"
  ddg.procedure(pname="subset variable",
                ins=list("all.data", "variable"),
                outs.snapshot=list("raw.data"))

  ddg.finish("read.data")
  return(raw.data)
}
```

Build into the workflows of research....

[Barbara S. Lerner and Emery R. Boose]
Components
Dependencies
Change

- 35 kinds of annotations
- 5 Main Workflows
- 14 Nested Workflows
- 25 Scripts
- 11 Configuration files
- 10 Software dependencies
- 1 Web Service
- Dataset: 90 galaxies observed in 3 bands
- Multiple platforms
- Multiple systems
You can download our code from the URL supplied. Good luck downloading the only postdoc who can get it to run, though #overlyhonestmethods

4:52 PM - 8 Jan 2013

313 RETWEETS 98 FAVORITES
Document vs Instrument

Reproducibility by Inspection
Read It

Reproducibility by Invocation
Run It
Instrument Entropy

all experiments become less reproducible

Mitigate
Detect, Repair
Preserve
Partial replication
Approx reproduce
Verification
Benchmarks

Zhao, Gomez-Perez, Belhajjame, Klyne, Garcia-Cuesta, Garrido, Hettne, Roos, De Roure and Goble. Why workflows break - Understanding and combating decay in Taverna workflows, 8th Intl Conf e-Science 2012
Workflow Planning & Watch of Workflows

Decay, Service Deprecation, Data source monitoring, Checklists, Minimal Models

Planning

Watch

monitor

monitor

monitor

Deploy

Operations

Repository

Workflows, myExperiment

Workflows for managing workflows
preservation

portability

packaging

provenance

gather dependencies

capture steps

track & keep results

versioning

variability tolerance

Portable Package

White Box, Installation
Archived record

Virtual Machines

Recompute, limited
installation, Black Box
Byte execution, copies

Open Source/Store

Descriptive read,
White Box
Archived record

Sci as a Service

Integrative fws

Read & Run, Co-location
No installation

[Adapted Freire, 2013]
preservation
portability
packaging

provenance
gather dependencies
capture steps
track & keep results

versioning
variability tolerance

Adapted Freire, 2013
Levels of Reproducibility

Portability

- Original Experiment
  - Binaries + Data
  - Source Code / Workflow + Data

- Similar Experiment
  - Binaries + Data + Dependencies
  - Source Code / Workflow + Data + Dependencies

- Different Experiment
  - Figures + Data
  - Binaries + Data + Dependencies
  - Virtual Machine
    - Source Code / Workflow + Data + Dependencies

Depth: how much of an experiment is available

Coverage: how much of an experiment is reproducible

[Freire, 2014]
FAIR

Findable
Accessible
Interoperable
Reusable

http://datafairport.org/
(Dynamic) Research Objects

- **Bundle** and **relate** multi-hosted digital resources of a scientific experiment or investigation using standard mechanisms, Currency of exchange

- Enabling **reproducible**, transparent research.

http://www.researchobject.org/
(Dynamic) Research Objects

- **Bundle** and **relate** multi-hosted digital resources of a scientific experiment or investigation using standard mechanisms, Currency.

- Enabling **reproducible**, transparent research.

---

http://www.researchobject.org/
(Dynamic) Research Objects

- **Bundle** and **relate** multi-hosted digital resources of a scientific experiment or investigation using standard mechanisms, Currency.

- Enabling **reproducible**, transparent research.

---

H. Van de Sompel et. al. **Persistent Identifiers for Scholarly Assets and the Web: The Need for an Unambiguous Mapping** 9th International Digital Curation Conference; **Trusty URLs**
Machine readable metadata*
Machine actionable systems**

* Especially Linked Data and RDF
** Especially REST APIs
Sys Bio Research Object

Systems Biology: Needed a common archive format for reuse across tools.

- Aggregation
- Annotations/provenance
- Ad-hoc domain-specific specification

OMEX archive

OMEX archive

OMEX archive

OMEX archive

OMEX archive

OMEX archive

OMEX archive

OMEX archive

OMEX archive

OMEX archive
The research lifecycle

IDEAS – HYPOTHESES – EXPERIMENTS – DATA - ANALYSIS - COMPREHENSION - DISSEMINATION

Commercial & Public Tools

Discipline-Based Metadata Standards

Git-like Resources By Discipline

Community Portals

New Reward Systems

Data Journals

Training

Institutional Repositories

Commercial Repositories

[Phil Bourne]
The Cameron Neylon Equation

\[
\text{Process} = \frac{\text{Interest}}{\text{Friction}} \times \text{Number people reach}
\]

Steps:
- Reproducible
- Born
- Towards

Productivity

Personal side effect

Public side effect

Reproducibility
“may all your problems be technical” ...Jim Gray

[Adapted, Daron Green]
Summary

- Workflow & modelling models in Science
- Software-style Stewardship
- Born reproducible
- Collective cost & responsibility
- Social factors dominate

http://www.force11.org

Force2015
12 - 13 January, 2015
Oxford University
• myGrid  
  – http://www.mygrid.org.uk
• Taverna  
  – http://www.taverna.org.uk
• myExperiment  
  – http://www.myexperiment.org
• BioCatalogue  
  – http://www.biocatalogue.org
• Biodiversity Catalogue  
  – http://www.biodiversitycatalogue.org
• Seek  
  – http://www.seek4science.org
• Rightfield  
  – http://www.rightfield.org.uk
• VPH-Share  
  – http://www.vph-share.eu/
• Wf4ever  
  – http://www wf4ever-project.org
• Software Sustainability Institute  
  – http://www.software.ac.uk
• BioVeL  
  – http://www.biovel.eu
• Force11  
  – http://www.force11.org
• SCAPE  
  – http://www.scape-project.eu/
Acknowledgements

- David De Roure
- Tim Clark
- Sean Bechhofer
- Robert Stevens
- Christine Borgman
- Victoria Stodden
- Marco Roos
- Jose Enrique Ruiz del Mazo
- Oscar Corcho
- Ian Cottam
- Steve Pettifer
- Magnus Rattray
- Chris Evelo
- Katy Wolstencroft
- Robin Williams
- Pinar Alper
- C. Titus Brown
- Greg Wilson
- Kristian Garza
- Donal Fellows
- Juliana Freire
- Jill Mesirov
- Simon Cockell
- Paolo Missier
- Paul Watson
- Gerhard Klimeck
- Matthias Obst
- Jun Zhao
- Pinar Alper
- Daniel Garijo
- Yolanda Gil
- James Taylor
- Alex Pico
- Sean Eddy
- Cameron Neylon
- Barend Mons
- Kristina Hettne
- Stian Soiland-Reyes
- Rebecca Lawrence
- Alan Williams

- Wf4ever, SysMO, BioVel, UTOPIA and myGrid teams