

# Research Data Management for Computational Science

Gerard Gorman  
g.gorman@imperial.ac.uk

Christian Jacobs  
c.jacobs10@imperial.ac.uk

# Data requirements

- Data produced by scientific software should be **reproducible** and **recomputable**.
- This requires:
  - **raw data** (input and output files)
  - the **software** (with info about the specific version used) to produce it
  - **provenance data**
- We need a way of **curating** this data and software at the push of a button...
- ...and a way of **referencing** it correctly in papers.

- In addition to papers and figures, **Figshare** ([figshare.com](https://figshare.com)) provides hosting for datasets.
- Each dataset is given its own **Digital Object Identifier (DOI)**.
- Programs developed by users can interface with Figshare via the Figshare API.

# Aims

- Develop a program to **automatically push software and data to Figshare**.
- Incorporate this program into the workflow of **Fluidity** – an open-source CFD code for fluid flow simulations (<http://amcg.es.ic.ac.uk/fluidity>).
- DOIs are ‘minted’ automatically, and added to the current **metadata** of simulation output.

# Fluidity with RDM support

## Current progress

- Implementation of a Python program which enables the publication of both software and data to Figshare.
- Addition of a 'publish' option to Fluidity simulation setup files.
- New DOIs created when:
  - Software is pushed to Figshare (if the specific version of the software, identified by the git commit hash, has not been published already).
  - Input data is pushed to Figshare.
  - Output data is pushed to Figshare.
- DOIs are recorded in the simulation setup file – if the simulation is run again, the same DOI is used to store the data.
- In the future, we will use MD5sums.

# Fluidity with RDM support

The screenshot shows a window titled "Diamond: top\_hat\_cg\_supg.flml (/home/christian/RDM/fluidity-rdm/tests/top\_hat\_cg\_supg)". The window contains a menu bar with "File", "Edit", "View", "Validate", "Tools", and "Help". Below the menu bar is a tree view of nodes. The "publish" node is highlighted in green and has a mouse cursor over it. To the right of the tree view is a panel with three sections: "Option Properties" (Description: No documentation), "Data" (No data), and "Comment" (No comment). The "publish" node has a red 'x' icon next to it, while other nodes have a plus sign icon.

Node	Icon
problem_type	
▶ geometry	
▶ io	
▶ timestepping	
▶ physical_parameters	✘
▶ material_phase (Fluid)	
material_phase	+
mesh_adaptivity	+
imported_solids	+
turbine_model	+
ocean_biology	+
ocean_forcing	+
reduced_model	+
porous_media	+
embedded_models	+
firedcomp	+
multiphase_interaction	+
▼ publish	✘
options_file	
▼ software	✘
article_id	✘
doi	✘
▶ input_data	✘
▶ output_data	✘

Screenshot of the 'publish' option in the Fluidity simulation setup file.

# Fluidity with RDM support

Example: simulation of the top\_hat\_cg\_supg test case

My data   Projects   Activity

11% of private storage used

<input type="checkbox"/>	<b>Add to Fileset</b> ▾ <b>Batch edit</b> ▾	Type ▾ mouseover(🇺🇸)
<input type="checkbox"/>	top_hat_cg_supg-output-data	DATASET
<input type="checkbox"/>	top_hat_cg_supg-input-data	DATASET
<input type="checkbox"/>	Fluidity-version-b68c9225ef2c84e827af39541bf45197d2468165	DATASET

Screenshot of software, input data and output data automatically pushed to Figshare.

- **The Software Sustainability Institute:** [www.software.ac.uk](http://www.software.ac.uk)
- **Digital Curation Centre:** [www.dcc.ac.uk](http://www.dcc.ac.uk)
- **Software Carpentry:** [software-carpentry.org](http://software-carpentry.org) (and **Data Carpentry:** [nescent.github.io/2014-05-08-datacarpentry](https://nescent.github.io/2014-05-08-datacarpentry))
- **Fidgit:** [www.github.com/arfon/fidgit](http://www.github.com/arfon/fidgit)
- **Reproducible Research** course:  
<https://www.coursera.org/course/repdata>
- **ROpenSci's Reproducibility Guide:**  
<http://ropensci.github.io/reproducibility-guide>

*'It has always been my habit to hide none of my methods, either from my friend Watson or from any one who might take an intelligent interest in them.'*

*Sherlock Holmes*