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.

Figshare

- In addition to papers and figures, Figshare (`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

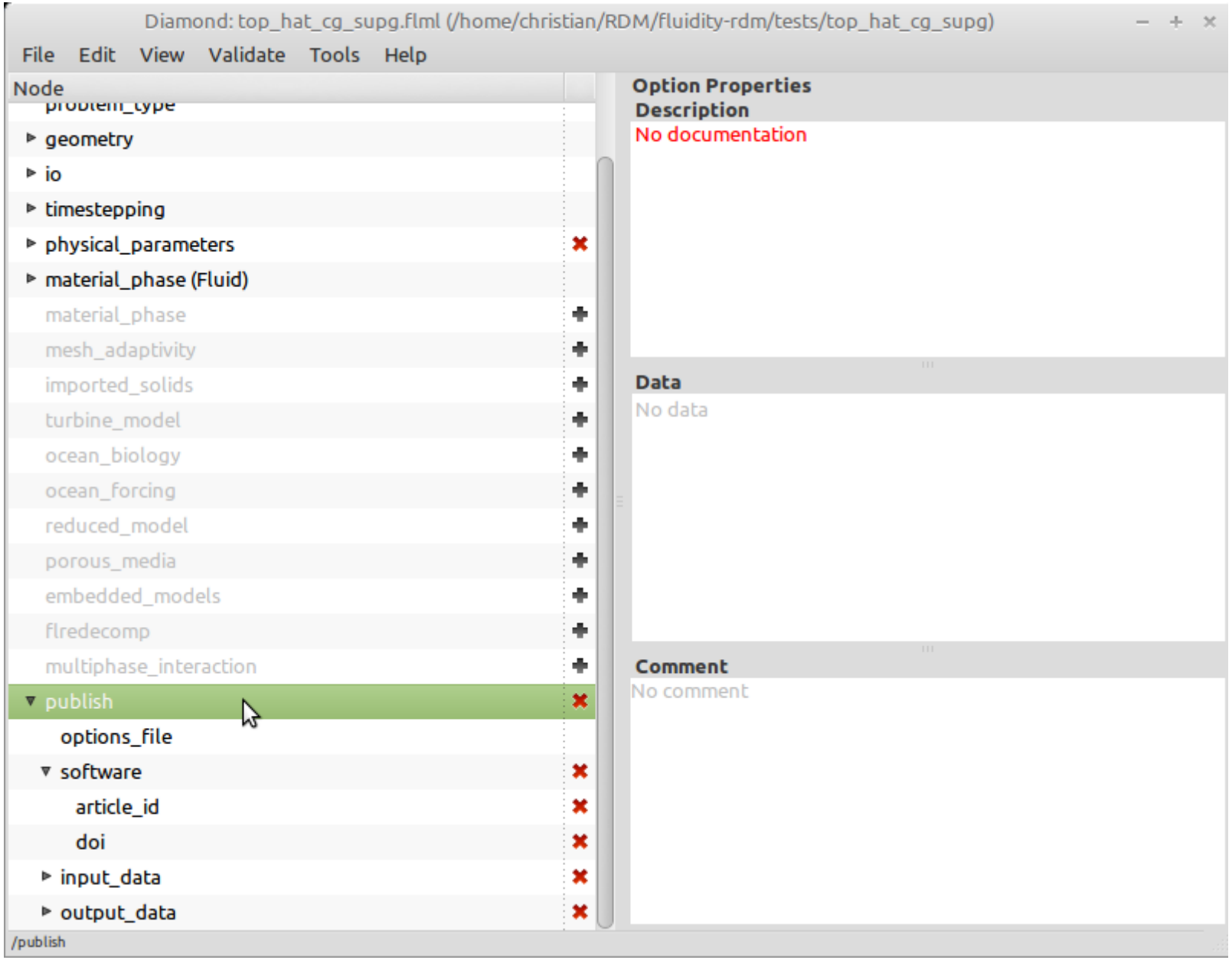
- Develop a program to automatically push software and data to Figshare.
- Incorporate this program into the workflow of Fluidity – an open-source CFD code for fluid flow simulations (`http://amcg.ese.ic.ac.uk/fluidity`).
- DOIs are 'minted' automatically, and added to the current metadata of simulation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

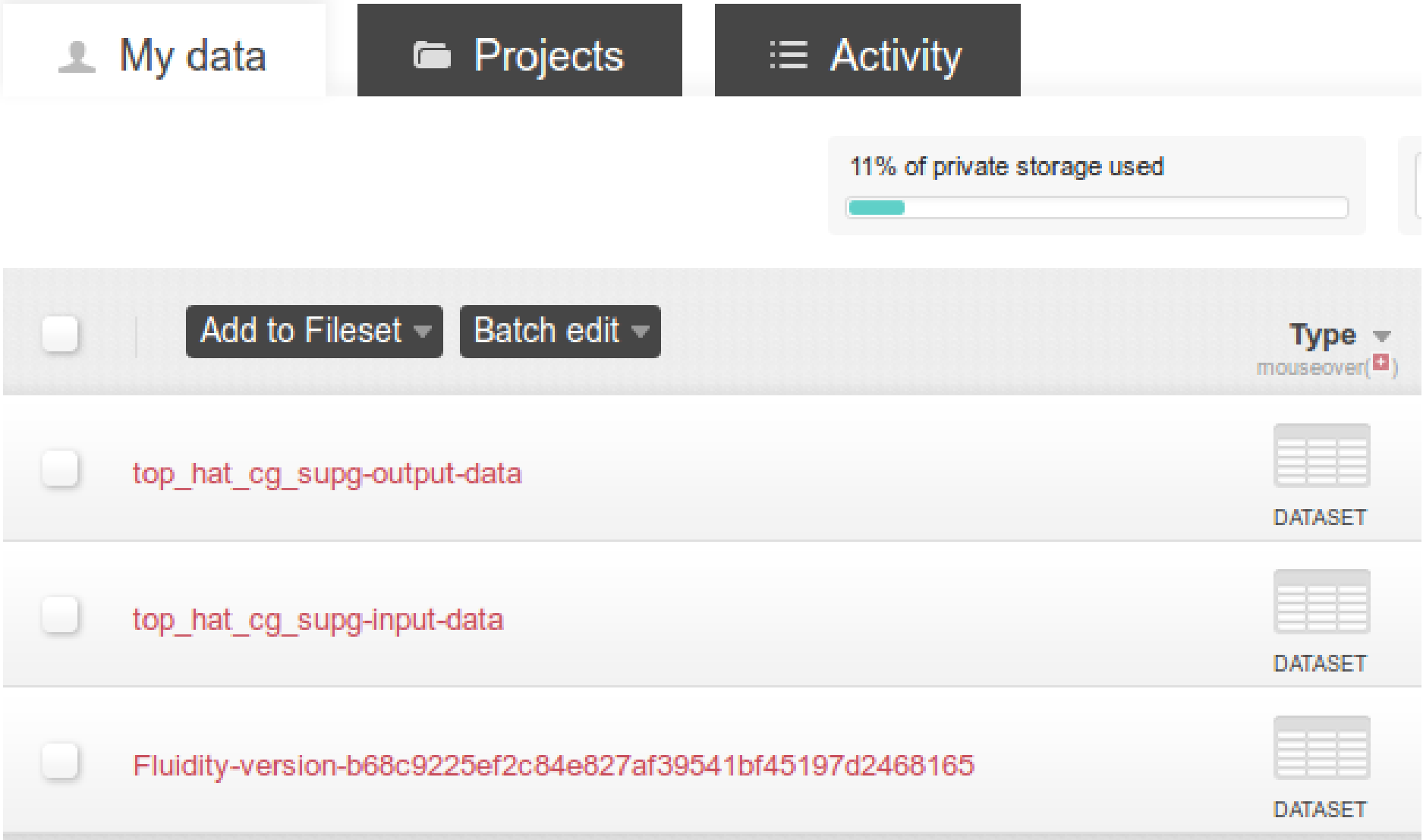
Fluidity with RDM support



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



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

Links

- The Software Sustainability Institute: `www.software.ac.uk`
- Digital Curation Centre: `www.dcc.ac.uk`
- Software Carpentry: `software-carpentry.org` (and Data Carpentry: `nescent.github.io/2014-05-08-datacarpentry`)
- Fidgit: `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