Research Data Management for Computational Science

Christian T. Jacobs[1]

c.jacobs10@imperial.ac.uk

www.christianjacobs.uk

@ctjacobs_uk

&

Alexandros Avdis[1], Simon L. Mouradian[1], Gerard J. Gorman[1], Matthew D. Piggott[1]

[1]Department of Earth Science and Engineering, Imperial College London

The Data Hide, ODSI, University of Sheffield

20 October 2015

Ocean Simulations

- Simulations of ocean dynamics are important in many applications.
  - Prediction of tsunami impacts

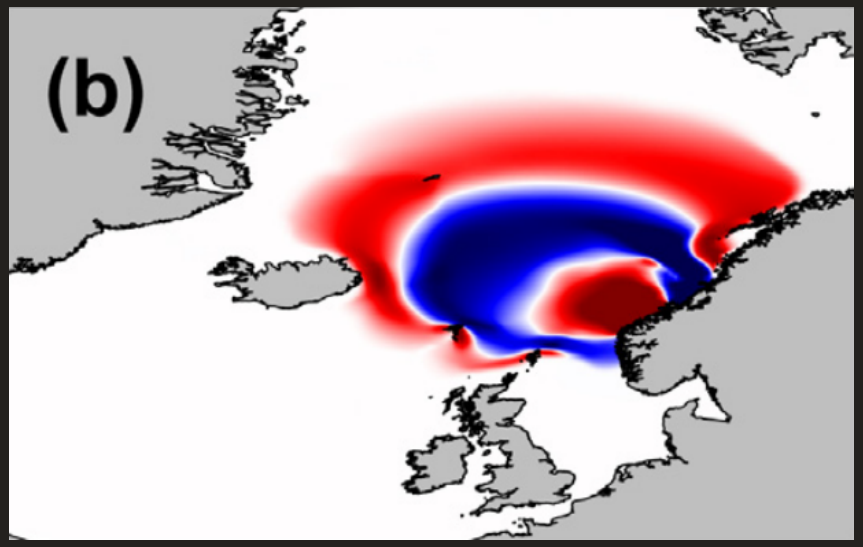


Image by Hill et al. (2014), used under CC-BY, doi:10.1016/j.ocemod.2014.08.007

  - Optimisation of marine renewable energy turbines
  - Estimating the range of nuclear contaminants

Software and Data Requirements

- Simulations should be recomputable and reproducible.
- This requires:
  - the software itself (with info about the specific version used)
  - raw data (input and output files)
  - provenance metadata

Problem

Unfortunately, most simulation-based publications are not accompanied by the data and the software (and exact version info) needed to recreate it.

What Can Be Done?

- The level of motivation amongst researchers to share their data and software is generally quite low.
  - Extra effort and time required to gather and publish it.
  - Typically gain little from the process.
  - See LeVeque et al. (2012)[1]

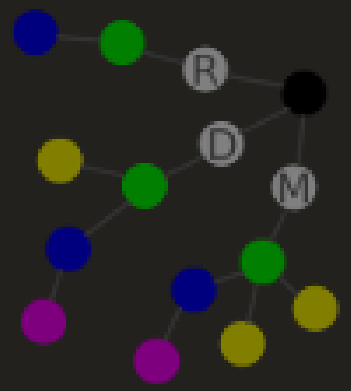
What we need

- We need a way of publishing data and software that is quick and easy...
- ...and a way of referencing it correctly in papers.

[1] LeVeque, R.J., Mitchell, I.M., Stodden, V. (2012). Reproducible Research for Scientific Computing: Tools and Strategies for Changing the Culture. Computing in Science & Engineering 14(4), 13--17.

"Green Shoots Project": PyRDM

- PyRDM: Research Data Management with Python
- Open-source, GNU GPL. `github.com/pyrdm/pyrdm`
- Facilitates the automated publication of source code and data to:
  - Figshare (`figshare.com`)
  - Zenodo (`zenodo.org`)
  - DSpace-based repositories (`dspace.org`)



Jacobs et al. (2014), DOI: 10.5334/jors.bj

- Online, citable and persistent repositories. Each code/dataset is given its own DOI.

Publishing Process: Software Source Code

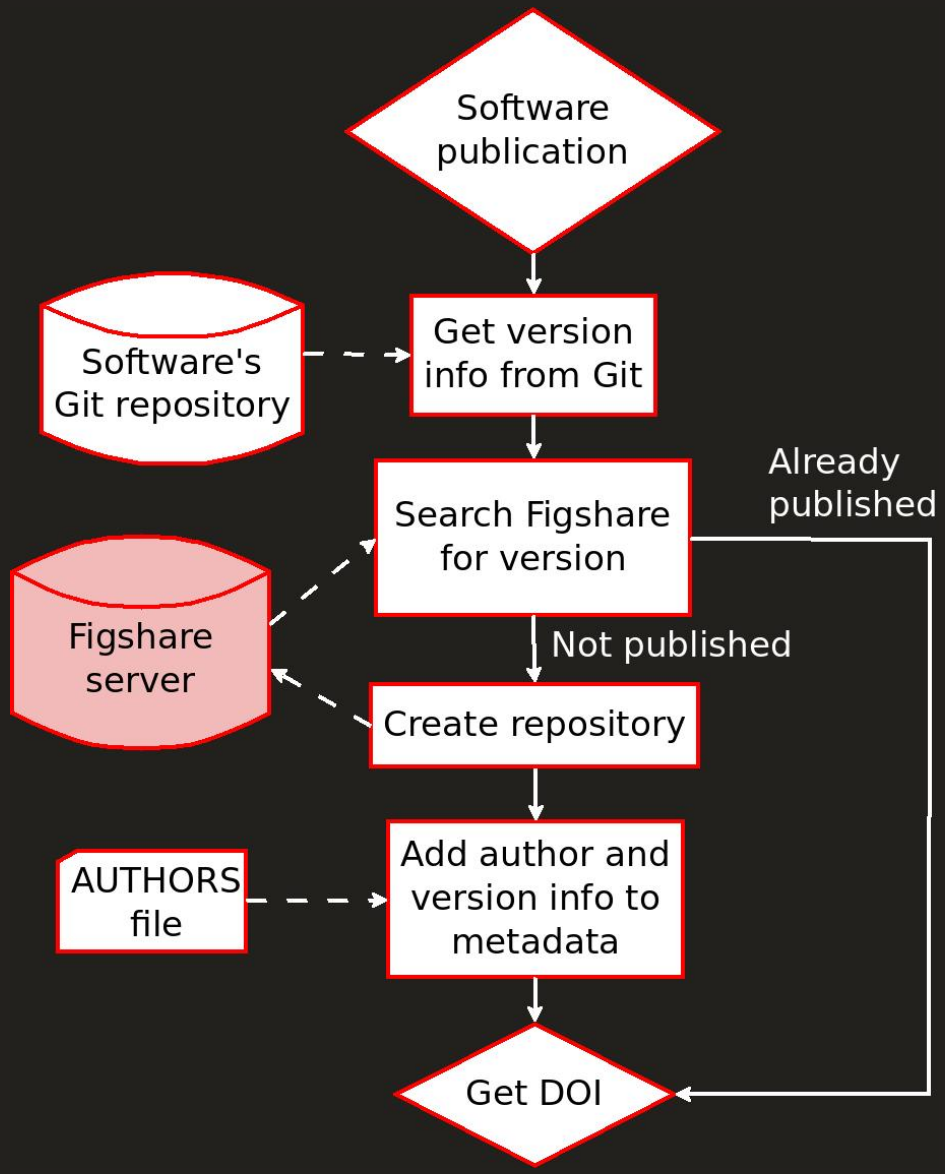


Image adapted from Jacobs et al. (2015).

Application to Ocean Simulations

- A prerequisite to a reproducible simulation is the availability and reproducibility of the mesh.
- Applied PyRDM to QMesh, a tool for generating meshes from GIS data (Avdis et al., in preparation).
  - See Jacobs et al. (2015) for details about RDM implementation.

Ocean simulations: The Mesh

- A key simulation input is the mesh.
  - Area of interest represented by discrete points/cells.

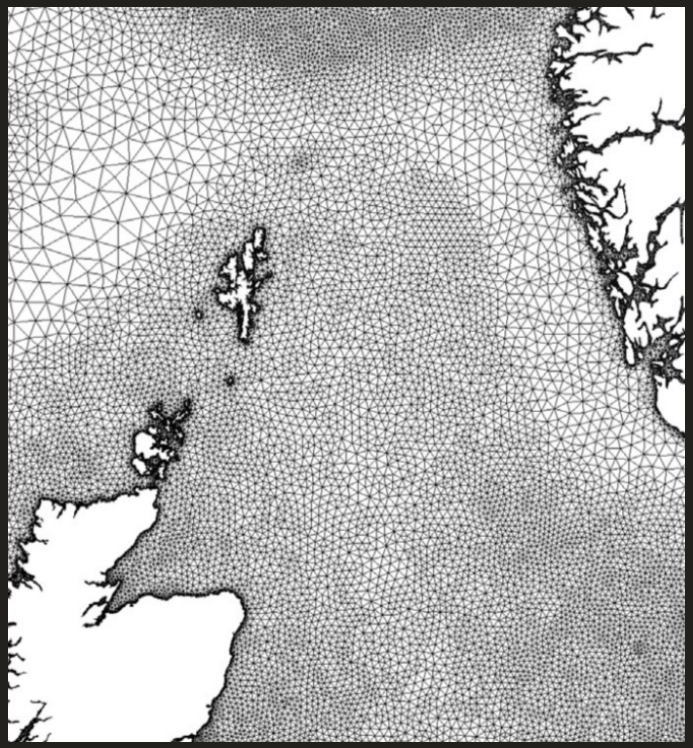


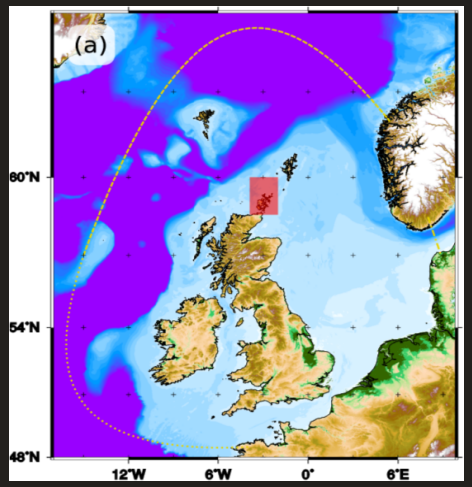
Image by Hill et al. (2014), used under CC-BY, doi:10.1016/j.ocemod.2014.08.007

- ...but creating a realistic, high-resolution mesh by hand is infeasible.

Geographical Information Systems

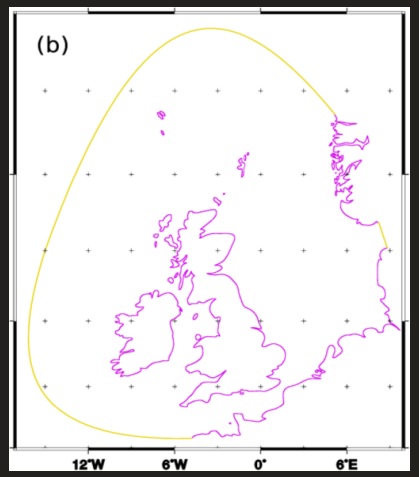
- Geographical Information Systems are good at processing bathymetry and coastline data to create a realistic geometry.
  - e.g. QGIS, ArcGIS, ...

Bathymetry data



+

Geometry



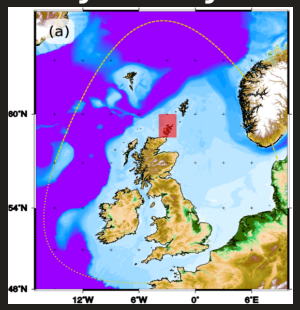
Images by Avdis et al. (2015).

- How do we create a mesh based on this input data?

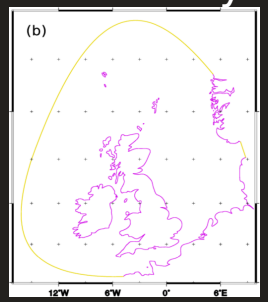
QMesh: Mesh Production using GIS Data

- QMesh is a software package which:
  - Takes the geometry defined in QGIS...
  - ...and converts the geometry into an appropriate format for...
  - ...Gmsh, a tool which generates the mesh for the domain.

Bathymetry data

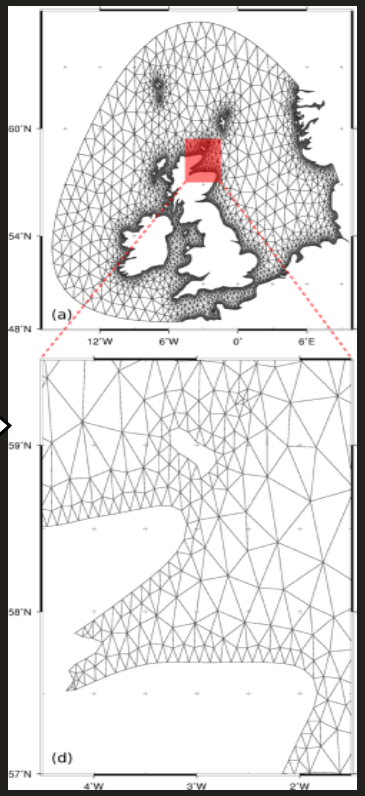


Geometry



QMesh converts to Gmsh format

Mesh



Images by Avdis et al. (2015).

Example Workflow: Orkney and Shetland Isles

- Consider the area around the Orkney and Shetland Isles.
- Involves a number of GIS input data files:
  - The QGIS project file itself, comprising:
  - Geometrical layer files defining the coastlines
  - Bathymetry data in a NetCDF file

Example Workflow: Geometry in QGIS

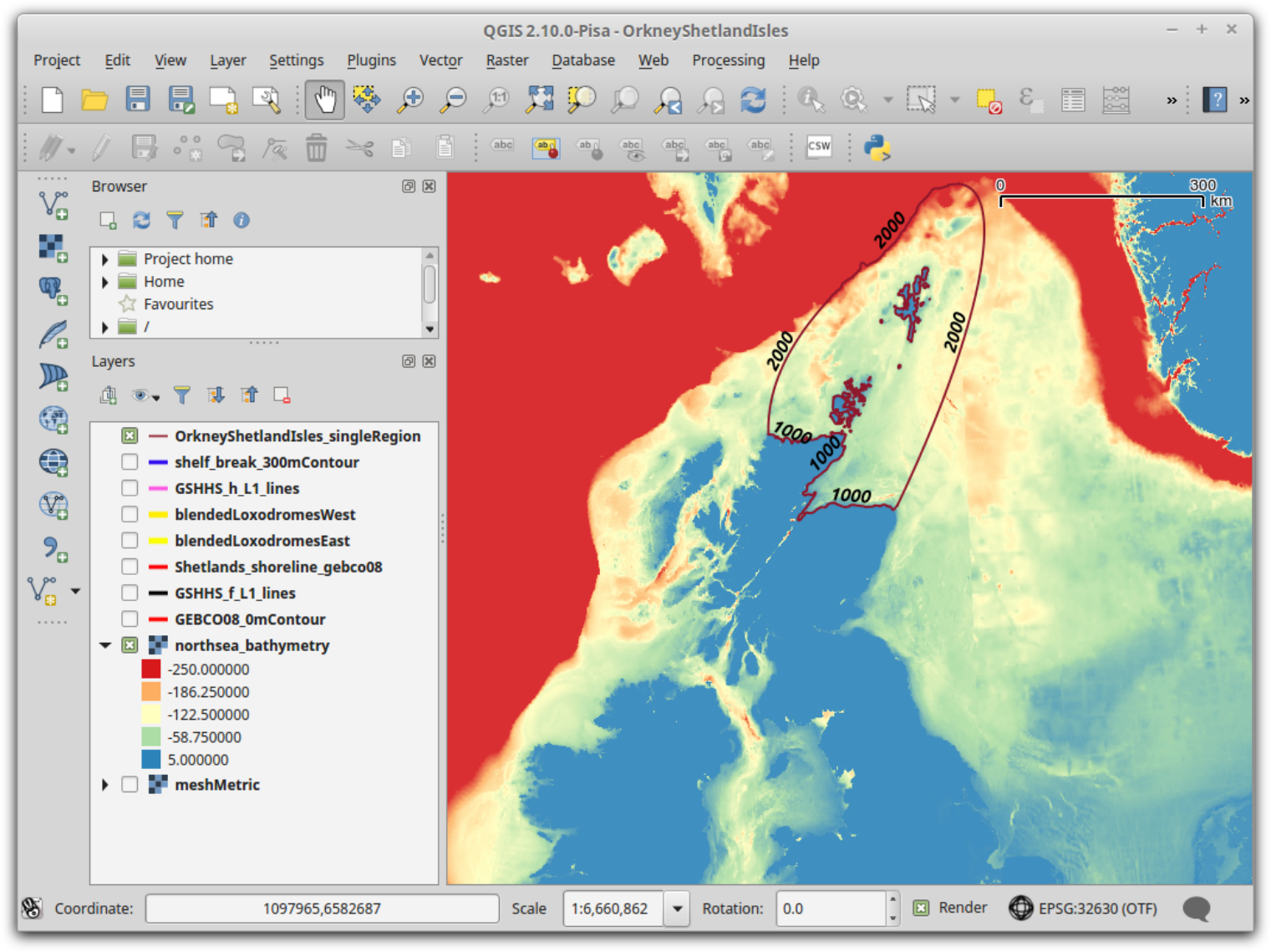


Image by Jacobs et al. (2015).

Example Workflow: Mesh from QMesh

- The input data in the QGIS project is used to produce a mesh using QMesh.
- User runs their ocean simulation using this mesh.
- When results are satisfactory, user publishes the data and software using the QMesh publishing tool.

Example Workflow: QMesh Publishing Tool

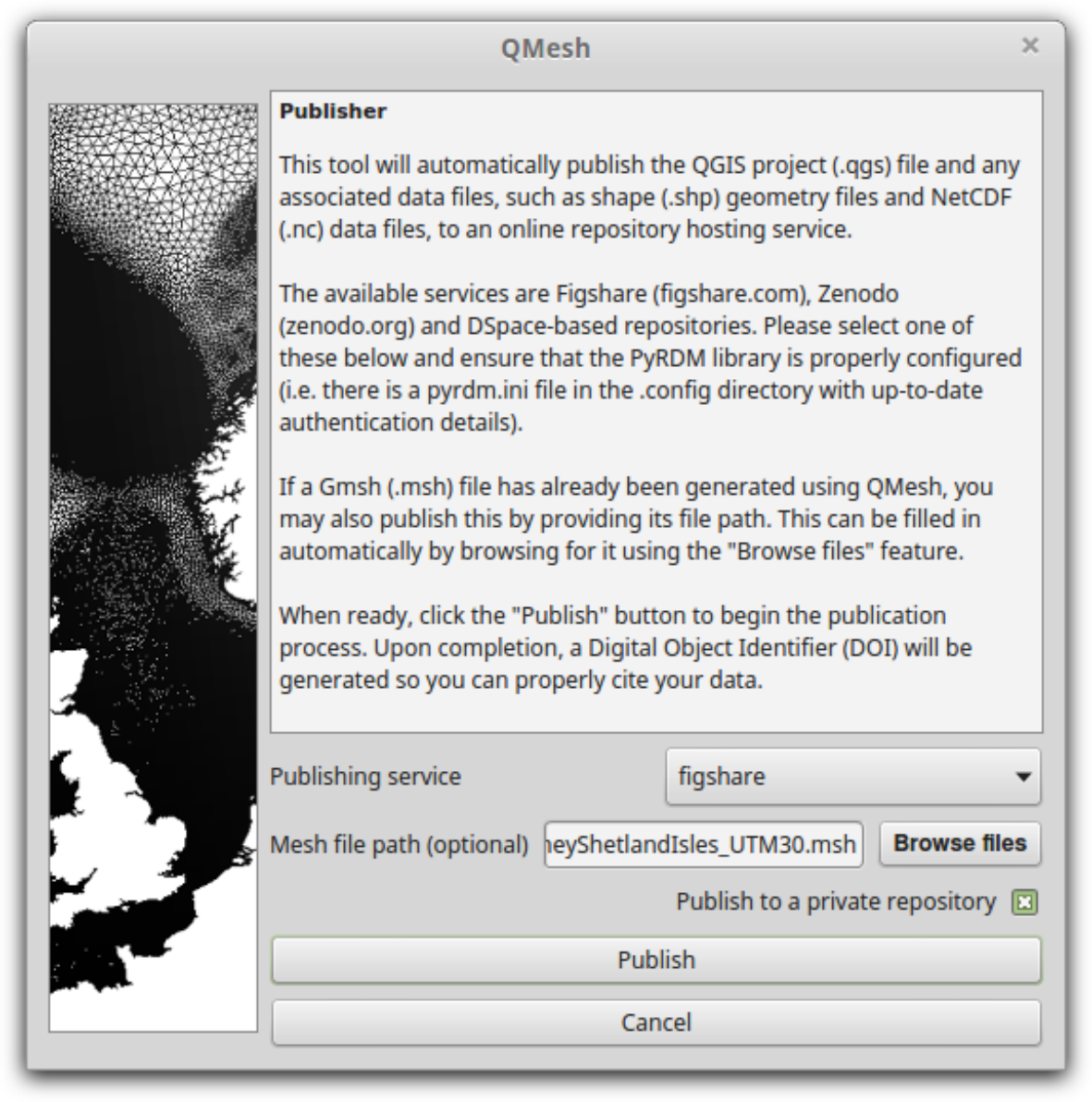


Image by Jacobs et al. (2015).

Publishing Process: Data

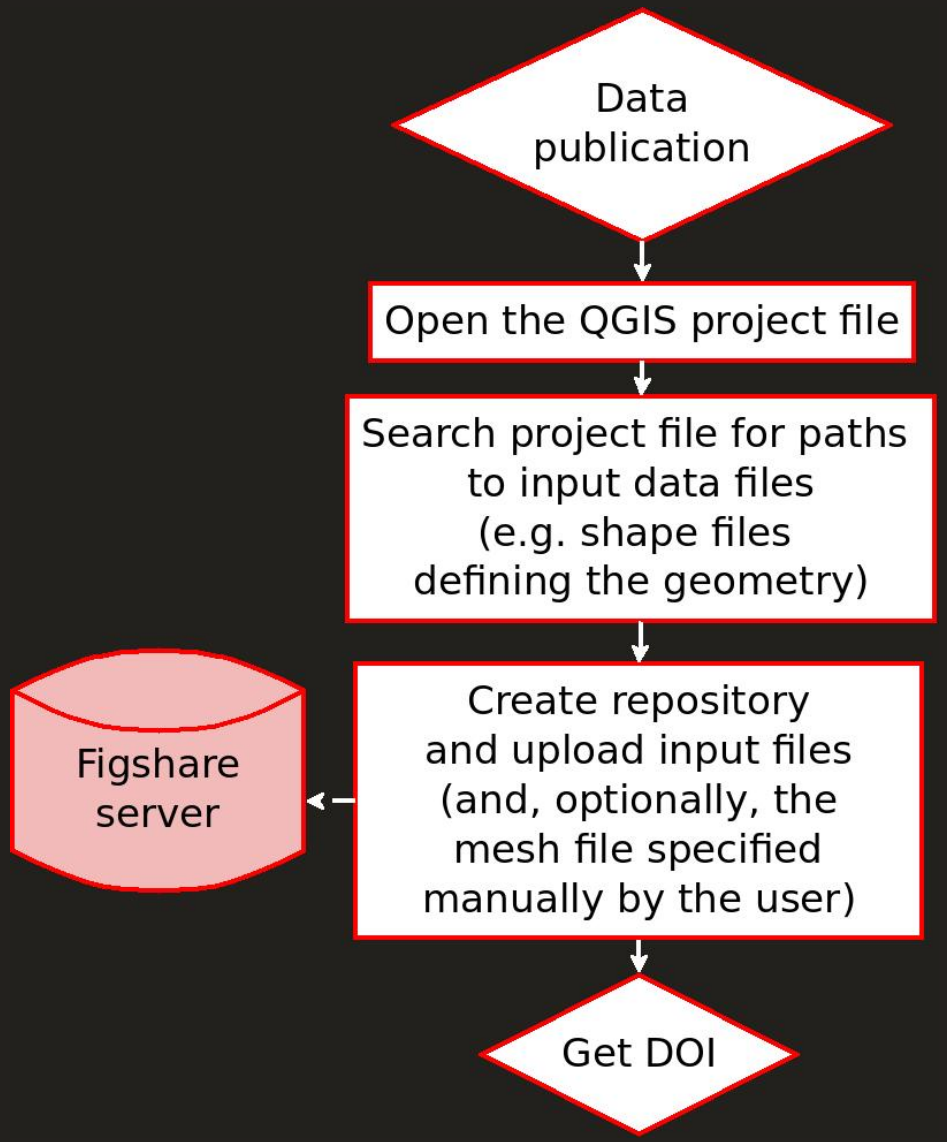


Image adapted from Jacobs et al. (2015).

Example Workflow: QGIS project file

- Publishing tool parses the XML-based QGIS project file to determine location of all data files that the project comprises...


```
      </edittypes>
    </maplayer>
    <maplayer minimumScale="0" maximumScale="1e+08" simplifyDrawingHints="1" minLabelScale="0" maxLabe
simplifyLocal="1" scaleBasedLabelVisibilityFlag="0">
      <id>Shetlands_shoreline_gebco08201407071159021 74</id>
      <datasource>./Shetlands_shoreline_gebco08.shp</datasource>
      <title></title>
      <abstract></abstract>
      <keywordList>
        <value></value>
      </keywordList>
      <layername>Shetlands_shoreline_gebco08</layername>
      <srs>
        <spatialrefsys>
          <proj4>+proj=longlat +datum=WGS84 +no_defs</proj4>
          <srsid>3452</srsid>
          <srid>4326</srid>
          <authid>EPSG:4326</authid>
          <description>WGS 84</description>
          <projectionacronym>longlat</projectionacronym>
          <ellipsoidacronym>WGS84</ellipsoidacronym>
          <geographicflag>true</geographicflag>
        </spatialrefsys>
      </srs>
```

Example Workflow: Files on Figshare

- ...and uploads these files to the repository hosting service via its API.

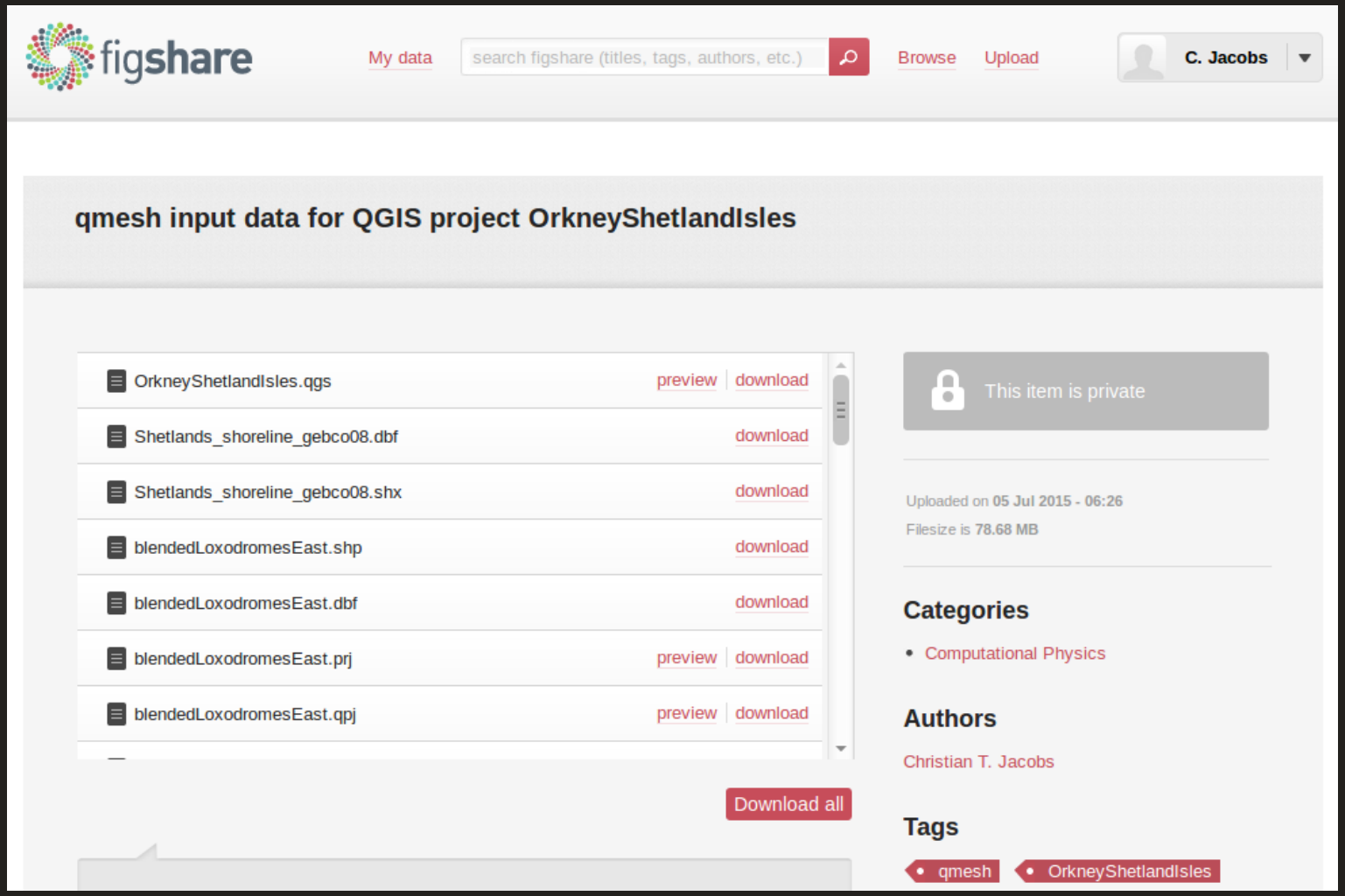


Image by Jacobs et al. (2015).

Example Workflow: DOI

Publication ID and DOI are assigned, and presented to user once publication process is complete:

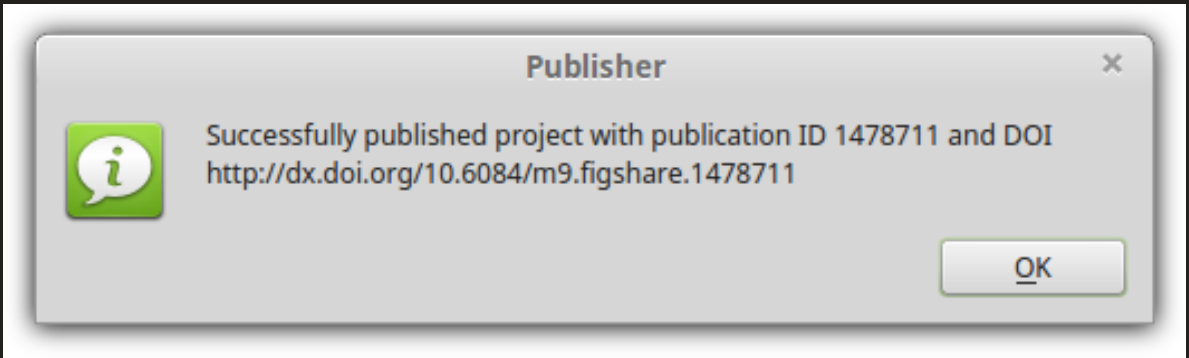


Image by Jacobs et al. (2015).

Issues/Limitations Encountered

- Lack of standardisation. Need a better way of affiliating authors.
- Lack of API support. No searching in Zenodo, no server-side MD5 checksums in Figshare, ...
- Restriction on private storage space.
- Restriction on number of collaborators.
- Figshare for Institutions / cloud storage to address these restrictions?
- Publishing QMesh source code may not be enough to reproduce the exact same mesh without knowledge of its dependencies.

References and Acknowledgements

- Jacobs et al. (2014). PyRDM: A Python-based library for automating the management and online publication of scientific software and data. Journal of Open Research Software, 2(1):e28. DOI: 10.5334/jors.bj
- Avdis et al. (2015). Shoreline and Bathymetry Approximation in Mesh Generation for Tidal Renewable Simulations. In Proceedings of the European Wave and Tidal Energy Conference (EWTEC) Series. Pre-print: http://arxiv.org/abs/1510.01560
- Avdis et al. (In Preparation). Efficient unstructured mesh generation for renewable tidal energy using Geographical Information Systems.
- Jacobs et al. (2015). Integrating Research Data Management into Geographical Information Systems. In Proceedings of the 5th International Workshop on Semantic Digital Archives. Pre-print: http://arxiv.org/abs/1509.04729

- Thanks to the Research Office at Imperial College London for funding.

- Slides produced using LaTeX, with a modified version of the Wronki Beamer theme (kaszkowiak.eu).

Research Data Management for Computational Science

Christian T. Jacobs[1]

c.jacobs10@imperial.ac.uk

www.christianjacobs.uk

@ctjacobs_uk

&

Alexandros Avdis[1], Simon L. Mouradian[1], Gerard J. Gorman[1], Matthew D. Piggott[1]

[1]Department of Earth Science and Engineering, Imperial College London

The Data Hide, ODSI, University of Sheffield

20 October 2015