

How to share your code, and why you should bother

Dr Lucy Whalley, October 2022

Hello!! 🖐️

2007-2011: Theoretical Physics MPhys, University of Birmingham

2010: Summer placement in gravitational waves group (my vague space science connection – also my introduction to **Python** and **Git**)

2011-2015: PGCE in post-compulsory education and training, primary school mathematics teacher

2015-2019: PhD in Materials Science, Imperial College London

2019: Fellowship with the **Software Sustainability Institute**

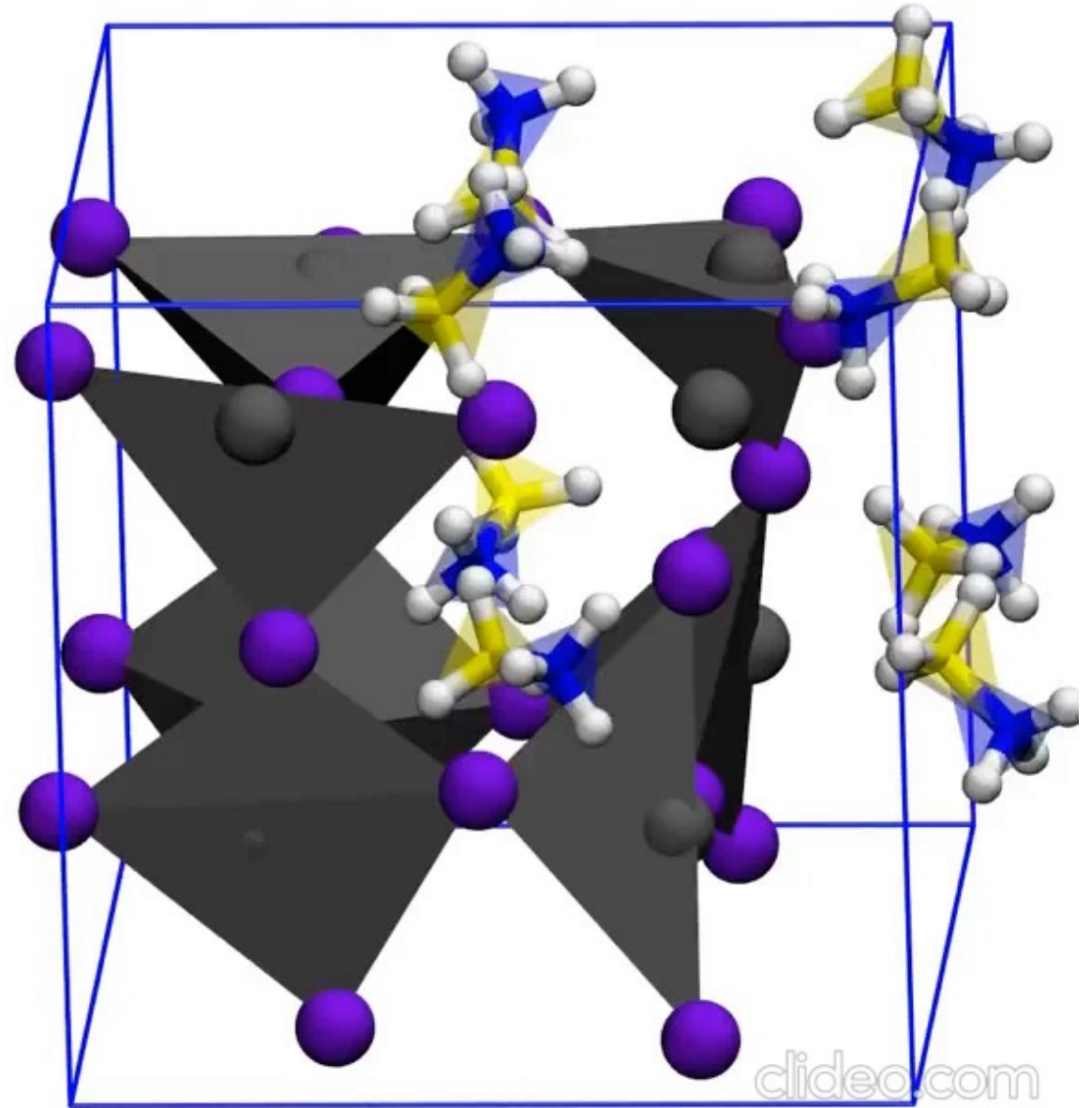
2020: Parental leave

2020: Joined Northumbria as a VCF

Hello!! 🖐️

Computational materials science:

solid state physics +
quantum chemistry +
high-performance-computing +
software engineering



Animation courtesy of Dr Jarvist Frost

Hello!! 🖐️

Computational materials science:

solid state physics +
quantum chemistry +

high-performance-computing +
software engineering

Accessibility
Reproducibility
Open [insert word here]
Impact
Publishing

Definitions: Reproducibility

From the US National Academies of Sciences, Engineering and Medicine:

Reproducibility

Obtaining consistent computational results using the same input data, computational steps, methods, code and conditions of analysis

Note: definitions differ – for an in-depth analysis of the different working definitions see “The fundamental principles of reproducibility”, Odd Erik Gundersen (RS Philosophical Transactions A, 2021)

Definitions: Open Source

OSI's Open Source Definition

- free redistribution
- source code availability
- derivatives allowed
- no limitations of who may use it or for what
- no additional license in place
- license must not depend on distribution format, technology, presence of other works



**open source
initiative®**

Key points

- Open code does not mean open source
- Open source requires you to provide an OSI-approved license
- There are options within OSI-approved licenses: e.g. – permissive (MIT) vs copyleft (GNU GPL)

<https://github.com/readme/guides/open-source-licensing>

<https://choosealicense.com>

Definitions: Repository

A software repository (“repo”) is a storage location for software packages

There are several
services available:



These are all based on the version control software Git:



First, I want to persuade you that sharing your code is a Good Idea. I'll start with a personal story...

Case Study: Effmass.py

- This was my first research software project
- `Effmass` calculates the effective mass of electrons in a particular material
- But the domain specific details aren't important..



Effmass circa 2016 🍲

- One module contains (long) methods for data parsing, analysis and print out
- In a private Github repo (with me as sole contributor)
- Has the functionality needed for my own data analysis
- No testing or documentation

Effmass circa 2018 🖥️

- Re-factored into six modules
- In a public Github repo
- Testing and continuous integration
- Documentation website

2019



Code published
With JOSS

Effmass now 💖

- 6 contributors
- Can parse data from multiple sources
- 14,000 downloads (PyPI stats), 18 citations

I didn't *know* other people would use effmass – so why did I put the time in?

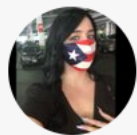
→ JOSS transformed the way I look at software

- It justified the time spent on learning new skills: documentation, testing, packaging (*I'll get a journal publication!*)
- The peer-review process forced me to share my code and this built my confidence (*I'm not the worst programmer in the world!*)



Some selfish reasons for sharing code

- If other people can use your code --> you can re-use your code --> more efficient working
- Valuable feedback through the peer review process
- Research credit through citation counts
- Career progression: the RSE career path, funding opportunities (e.g. EPSRC)
- Appreciation from your colleagues and self-promotion



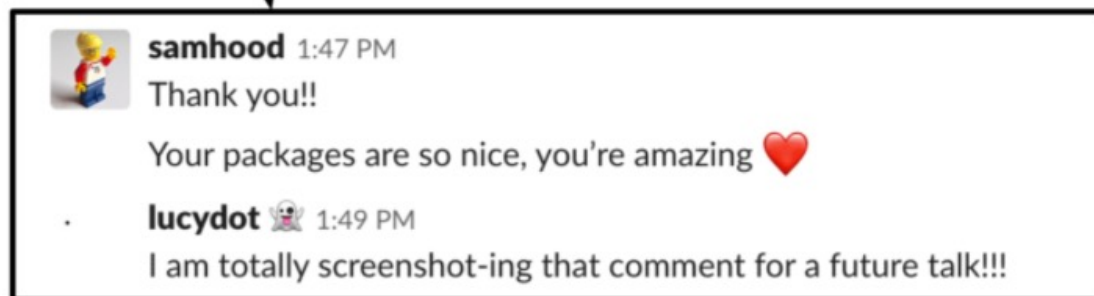
Giannina Guzmán Caloca 🇵🇷 @GianniG97 · 5 Aug



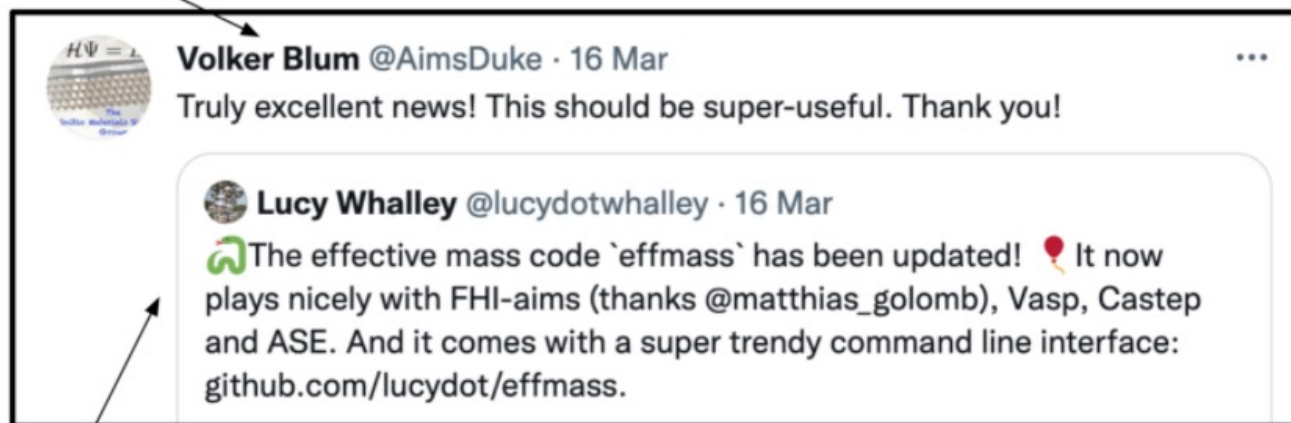
Hey Astro twitter, it is me once again here to make my yearly reminder thread about how software development projects are NOT respected or celebrated in this field as much as they should be despite carrying a lot of research on their backs and why/how that affects all of us (1/n)



Happy colleague



Big person in the field



Little person in the field

What about the not-so-selfish reasons?

- Other people can use your code and the field will progress more rapidly
- To ensure scientific reproducibility

We have publication processes to root out error for research that is done without a computer. Once you introduce a computer, the materials section in a typical scientific paper doesn't come close to providing the information that you need to verify the results. Analysing complicated data by computer requires instructions consisting of script and code. Hence we need the code, and we need the data.

Victoria Stodden, Editor, Journal of the American Statistical Association

Computers let you make mistakes faster than any invention in human history

- Odd Erik Gundersen, "The fundamental principles of reproducibility"

What are my code sharing options?

Public repo + citation file:

- ✓ Straight forward to implement
- ✗ No peer review

What is a CITATION.cff file?

`CITATION.cff` files are plain text files with human- and machine-readable citation information for software (and datasets). Code developers can include them in their repositories to let others know how to correctly cite their software.

This is an example of a simple `CITATION.cff` file:

```
cff-version: 1.2.0
message: "If you use this software, please cite it as below."
authors:
  - family-names: Druskat
    given-names: Stephan
    orcid: https://orcid.org/0000-0003-4925-7248
title: "My Research Software"
version: 2.0.4
```



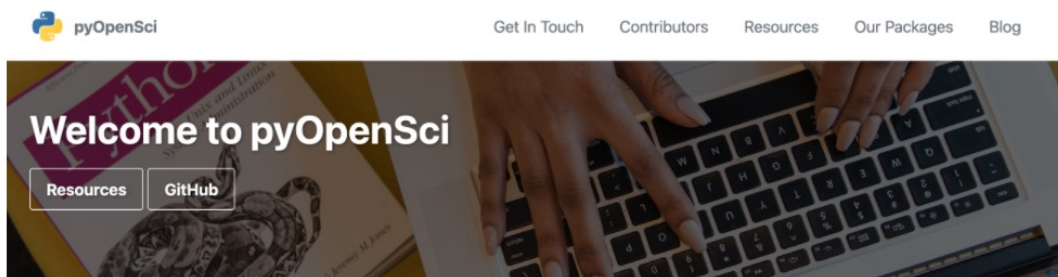
Arfon Smith ✓ @arfon · 16 Sep

In the same period, more than 17,000 CITATION.cff files have been pushed to GitHub, including some high profile projects such as [github.com/tensorflow/ten...](#), [github.com/huggingface/tr...](#), [github.com/JuliaLang/julia](#), [github.com/pandas-dev/pan...](#), [github.com/borisdayma/dal...](#) and a whole bunch more 😊

What are my code sharing options?

Code review with a colleague or community member:

- ✓ A good option if you are nervous about releasing your code to the wild!
- ✓ Several on-line initiatives if there is no-one in your immediate research circle
- ✗ No citation



About pyOpenSci

pyOpenSci promotes open and reproducible research through peer-review of scientific Python packages. We also build technical capacity by providing a curated repository of high-quality packages and enabling scientists to write and share their own software. We hope to foster a greater sense of community among scientific Python users so that we can help each other become better programmers and researchers. See our [list of Python packages](#) for an idea of the open-source projects that pyOpenSci has assisted.

pyOpenSci is being modeled after the successful [rOpenSci](#) community.



About Programmes and Events Resources

Apply to our Code Review Clinic

Posted by j.laird on 18 October 2021 - 9:46am

As part of our upcoming Research Software Camp: Beyond the Spreadsheet we will be running a Code Review Clinic. This will allow first time/beginner coders to get feedback on their code from an expert.

If you've recently started writing your own code related to research and are looking for some feedback on your code then our Code Review Clinic is for you! The Code Review Clinic will run on week 2 (8-12 November 2021) of the **Research Software Camp: Beyond the Spreadsheet**. There will be one session in the morning and one in the afternoon (both GMT) from Monday to Friday.

In the application form, you'll be asked to describe the code and project (if available) you want to submit for review. We welcome applications from first time/beginner coders mostly, but we're accepting applications from other levels of expertise too. This is meant to be a welcoming space for beginner coders, and, as such, we welcome code at any stage of its development.

Visit the **Code Review Clinic** page for further information.

[Apply to the Code Review Clinic](#)

Applications close on Wednesday 27 October.



Photo by [Agence Olloweb](#)

Tags

- [Jacalyn Laird](#)
- [Research Software Camps](#)
- [Communications](#)

What are my code sharing options?

Executable paper (e.g. a Jupyter Notebook) as Supplementary Information:

- ✓ Citeable
- ✗ Code is usually not peer-reviewed
- ✗ Limited to smaller pieces of code
- ✗ Requires a corresponding full length article

To enable notebook interaction
without the user having to install
anything



master MAPI_iodine_vibrations / IPR_data / IPR_analysis.ipynb Go to file ...

lucydot updated IPR analysis Latest commit defa62e on 15 Feb 2021 History

1 contributor

623 lines (623 sloc) 53.5 KB

```
In [1]: import pickle
import matplotlib.pyplot as plt
import itertools
%matplotlib inline
```

Introduction

- This notebook contains steps to analyse the harmonic phonon properties of the hybrid halide perovskite MAPb with a negatively charged iodine interstitial defect
- IPR_dictionary.p contains various properties (including inverse participation ratio) of each gamma-point harmonic phonon mode in a 222 supercell (96-atom) of MAPb with a negative iodine interstitial.
- The defect structure is included in this folder as POSCAR.vasp

Associated literature

Further details (including structural relaxation and phonon calculation details) can be found in:

- L. D. Whalley et al., 2020. *Giant Huang-Rhys factor for electron capture at the iodine interstitial in MAPb*
- L.D. Whalley, 2019. PhD thesis, *Defects and Distortions in Hybrid Halide Perovskites* (Chapter 6)

Inverse Participation Ratio

To identify which phonon modes are associated with the iodine interstitial defect we use the Inverse Participation Ratio (IPR) to quantify localisation.

The IPR is calculated from the harmonic phonon eigenvectors e_i of the system and is given by

Publishing code in a traditional journal

Publishing in a traditional computational journal (e.g. Journal of computational electronics)

- ✓ Citeable
- * Code is usually not necessarily peer-reviewed
- * Requires mapping your code to a written journal article (*is this a good use of resources?*)

"...the basic means of communicating scientific results hasn't changed for 400 years. Papers may be posted online, but they're still text and pictures on a page."

From *The Scientific Paper Is Obsolete* by James Somers



SPACE: 3D parallel solvers for Vlasov-Maxwell and Vlasov-Poisson equations for relativistic plasmas with atomic transformations ☆

Kwangmin Yu ^a, Prabhat Kumar ^{b, 1}, Shaohua Yuan ^b, Aiqi Cheng ^b, Roman Samulyak ^{b, a} ✉

[Show more](#) ✓

[+](#) Add to Mendeley [🔗](#) Share [🗣️](#) Cite



Sarkas: A fast pure-python molecular dynamics suite for plasma physics ☆, ☆☆

Luciano G. Silvestri ^a ✉, Lucas J. Stanek ^a, Gautham Dharuman ^b, Yongjun Choi ^a, Michael S. Murillo ^a

[Show more](#) ✓

[+](#) Add to Mendeley [🔗](#) Share [🗣️](#) Cite

<https://doi.org/10.1016/j.cpc.2021.108245>

[Get rights and content](#)

Publishing code in a developer friendly journal

Publishing in a developer friendly journal (e.g. The Journal of Open Source Software, The Journal of Open Research Software)

- ✓ A citeable journal publication
- ✓ Code is peer-reviewed
- ✓ Time efficient: The paper can be prepared in less than an hour



Dan Foreman-Mackey (@dfm)

: Astrophysics, probabilistic programming, data science, python

Dan Foreman-Mackey is a Research Scientist at the Flatiron Institute in the Center for Computational Astrophysics. His research program focuses on the development and application of probabilistic data analysis techniques to make novel discoveries and solve fundamental problems in astrophysics.

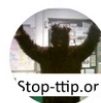


Monica Bobra (@mbobra)

Editor: Data Science, Heliophysics, Space Weather

Monica Bobra works as a Senior Research Data Scientist at Tomorrow.io, where she predicts terrestrial weather. She previously worked on predicting space weather at Stanford University and the Harvard-Smithsonian Center for Astrophysics.

JOSS is free and uses a fully transparent peer review process



Peter Murray-Rust @petermurrayrust · 8 Jul

6/ I emphasize that JOSS is *top class*. Just because there is no price to readers or authors does not affect quality. I would regard software published in JOSS as at least the equal of and probably better than published in a "high-impact" journal.



SunPy: A Python package for Solar Physics

Stuart J. Mumford^{1, 2, 3}, Nabil Freij⁴, Steven Christe⁵, Jack Ireland⁵, Florian Mayer⁶, V. Keith Hughitt⁷, Albert Y. Shih⁵, Daniel F. Ryan^{8, 5}, Simon Liedtke⁶, David Pérez-Suárez⁹, Pritish Chakraborty¹⁰, Vishnunarayan K I.⁶, Andrew Inglis¹¹, Punyaslok Pattnaik¹², Brigitta Sipőcz¹³, Rishabh Sharma⁶, Andrew Leonard³, David Stansby¹⁴, Russell Hewett¹⁵, Alex Hamilton⁶, Laura Hayes⁵, Asish Panda⁶, Matt Earnshaw⁶, Nitin Choudhary¹⁶, Ankit Kumar⁶, Prateek Chanda¹⁷, Md Akramul Haque¹⁸, Michael S Kirk¹¹, Michael Mueller⁶, Sudarshan Konge⁶, Rajul Srivastava⁶, Yash Jain¹⁹, Samuel Bennett⁶, Ankit Baruah⁶, Will Barnes²⁰, Michael Charlton⁶, Shane Maloney²¹, Nicky Chorley²², Himanshu⁶, Sanskar Modi⁶, James Paul Mason⁶, Naman9639⁶, Jose Ivan Campos Rozo²³, Larry Manley⁶, Agneet Chatterjee²⁴, John Evans⁶, Michael Malocha⁶, Monica G. Bobra²⁵, Sourav Ghosh²⁴, Airmansmith97⁶, Dominik Stańczak²⁶, Ruben De

A summary of our code sharing options

Publishing method	Example	Citation?	Software peer-review?	Journal publication?	Time-efficient?
Public repo + citation file	Citation File Format	✓	✗	✗	✓
Community peer review	rOpenSci, pyOpenSci	✓	✓	✗	✓
Executable paper as Supplementary Information	Jupyter Notebook	✓	✗	✓	✓
Software paper	Journal of Computational Electronics	✓	✗	✓	✗
Software meta-paper	JOSS, JORS	✓	✓	✓	✓

List of software journals: <https://www.software.ac.uk/which-journals-should-i-publish-my-software>

But my code isn't good enough to share

Yes it is. It doesn't need to be perfect. Sharing your poorly documented, untested, messy code is better than sharing no code. If you want to see an example of bad code that is being shared publicly, feel free to visit my Github (username: lucydot) 🤪

Tools

- **Github** (for development, testing, distribution)
 - **Github Actions** allows automatic testing
 - Can be used to build/host documentation: <https://nucem.github.io/ThermoPot>
 - Also useful for teaching: https://lucydot.github.io/python_novice/
 - **Gitlab** has similar functionality
- Mkdocs or ReadTheDocs for documentation websites
 - Automatic API documentation from docstrings
- Simple testing in Python – pytest
- Binder for interactive Notebooks
- Docker for reproducible environments

Communities

- Software Sustainability Institute (they have a Fellowship scheme,)
- Society of Research Software Engineering (who have a friendly annual conference)
- Software Carpentry (if you are learning or teaching basic computational skills)
- JOSS (Journal of Open Source Software)
- JORS (Journal of Open Research Software)
- ReproHacks – reprohack.org
- Northumbria research computing community (see the Teams site)
- Watch out for: Research Software Engineering Northern group

The rise of the RSEs – Research Software Engineer(ing)

RESEARCH SOFTWARE ENGINEERS - STATE OF THE NATION REPORT - 2017

Most research would be impossible without software, and this reliance is forcing a rethink of the skills needed in a traditional research group. With the emergence of software as the pre-eminent research tool used across all disciplines, comes the realisation that a significant majority of results are based, ultimately, on the skill of the experts who design and build that software. The UK has led the world in supporting a new role in academia: the Research Software Engineer (RSE).



Society of Research Software Engineering

Research Software Engineer – Social Sciences & Humanities

Published	Deadline	Location
28 Sep	10 Oct	Amsterdam

Demographics

What is your highest level of education?



● Doctorate: 70%
● Masters: 18%
● Undergraduate: 12%

Has your software contributed to a publication?



● Yes: 88%
● No: 12%

Ok, you've heard me speak.....now I have some questions!

- Do you use any of these tools?
- Which tools do you use for sharing code and data? Testing code? Documenting code and data?
- How would these ideas translate to data-driven science (Big Data) and Machine Learning?
- What do you feel is the culture in space and solar physics around sharing code and data?

Email me: l.whalley@northumbria.ac.uk

Slides will be on my website: lucydot.github.io/talks