Reproducibility and code sharing: IPR and risk assessment

Louise Corti

Head Development and Impact, Secure Research Service

Office for National Statistics, UK

8 June 2021

Office for National Statistics

Would you share your last Rolo?

Through engagements, negotiations with data owners and researchers around sharing data, documentation, code:

- Honing benefits' arguments and persuasive skills
- Helping build cultures
- Writing on arguments and strategies

Extremes in perceptions and reactions:

- Eager sharers, early adopters, creative strategies, champions
- Want to, concerns, trouble 'letting go'
- It's mine, I won't /can't share it

Social surveys: what code is already available?

Despite research transparency drivers, very little is published in the public domain:

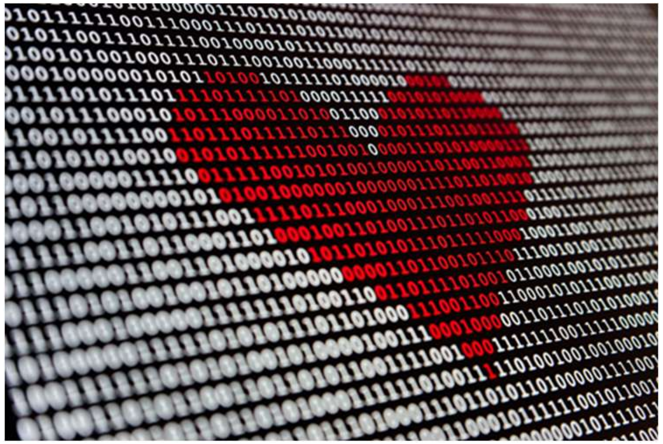
- **Data owners** - could make their data reproducible when generating 'research-ready' datasets
    - Documenting the full provenance chain (creation, cleaning, versioning, DVs, usability)

- **Researchers, peer community and higher educational institutions** – can strive to follow best practice on being reproducible
    - Willingness to learn new skills e.g. good coding, use of code tracking software, encouraging capacity building, understanding and asserting rights in the code

Data owners

Most survey data producers do not reveal code for processing operations or derived variables:

- Dedicated session at February 2020 #LoveYourCode event to hear views
- Issues around and intentions on publishing own code
- Thoughts about user-generated code and its status



Data owners: ONS

- For surveys, DV code is rarely part of the package of user documentation (detailed technical report/code books)
- LFS produced a detailed report on workflows that show the origins of standard DVs, but underlying code is not made available, or on request
- Emerging standards and frameworks across government that will help to support trust in data and the statistical production chain:
  - Reproducible pipelines (transformation, cleaning and analysis operations) increasingly being as reusable code
  - Standards, assessing compliance, trustworthiness, quality and value (TQV): Code of Practice for Statistics
  - Established Quality Assurance Toolkits e.g. Administrative Data (QAAD)

Data owners: NatCen

- Health Survey for England/British Social Attitudes Survey; create @250-800 DVs per survey release

- Use in house protocols, using comprehensive code and second checking, consistent variable naming, short human description of what the derived variable aims to do

- Desire to publish in-house generated supporting code

Thanks to Jess Bailey

Data owners: UCL CLS Birth Cohort Studies

- Data managers follow a protocol for creating detailed logic and algorithms for syntax that includes exact variable names, values etc.
- Histories are complicated (loops, arrays, macros, etc.)
- User Guides show how derived variables were created, published as user documentation
- Plans to release own code via GitHub

Thanks to Aida Sanchez

Cohort and Longitudinal Studies Enhancement Resources (CLOSER)

- Efforts to document code for creating harmonised variables (post-hoc)

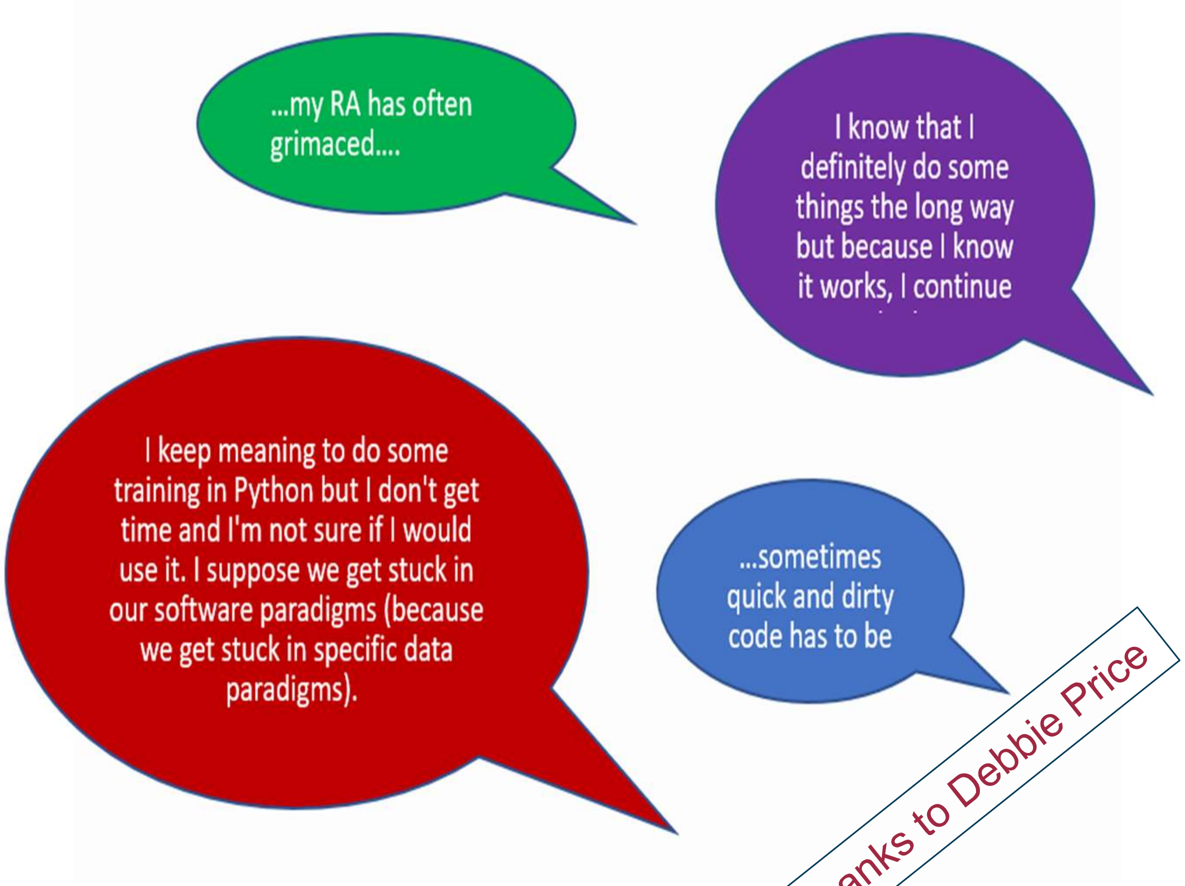
- Systematic approach to harmonising:
  - agreed code style
  - standardised documentation and metadata templates
  - link code directly to the published metadata available in CLOSER Discovery

- Difficulties arise for undocumented historical data collection and management

Thanks to Dara O'Neill

User generated code: current practices

- Not shared
- Available on request
- Submitted to journal as supplementary material
- Self-published on GitHub
- Published in community GitHub
- Formally published in a repository with a DOI


...my RA has often grimaced....
I know that I definitely do some things the long way but because I know it works, I continue
I keep meaning to do some training in Python but I don't get time and I'm not sure if I would use it. I suppose we get stuck in our software paradigms (because we get stuck in specific data paradigms).
...sometimes quick and dirty code has to be

Thanks to Debbie Price

Data owners: user generated code

- Few survey data owners have actively considered what to do with user generated code

- What do they think about user-created code?
  - Overall lack of resource to QA other people's new code
  - Might poor quality be a risk to the survey's reputation?
  - Often don’t get to know about new added value code out there
  - Who owns that code?
  - Who should get credit for the code?

Summary: reputational issues with publishing code

- Who 'owns' it?

- How to cite it? Who gets credit?

- Quality assessment – whose responsibility?

  Just because its reproducible, doesn’t mean its good quality…

- Risk assessment - whose responsibility?

Code IPR and licensing statements

✓ Decide where best to publish the code
  - Data owner documentation?
  - Code repository or GitHub; closed area for 'sensitive code'
  - Repository record, e.g. Institutional repository

✓ Agree QA and onward sharing licence
  - Validation/assessment protocols
  - Reproducibility 'certified'?
  - Disclaimers

✓ Agree and declare code 'authors'

✓ Declare/reference any original data sources
  - Ideally with DOIs with a public landing page
  - Helps with journal's Data Availability statements

VALIDATED

Summary: Owner's code

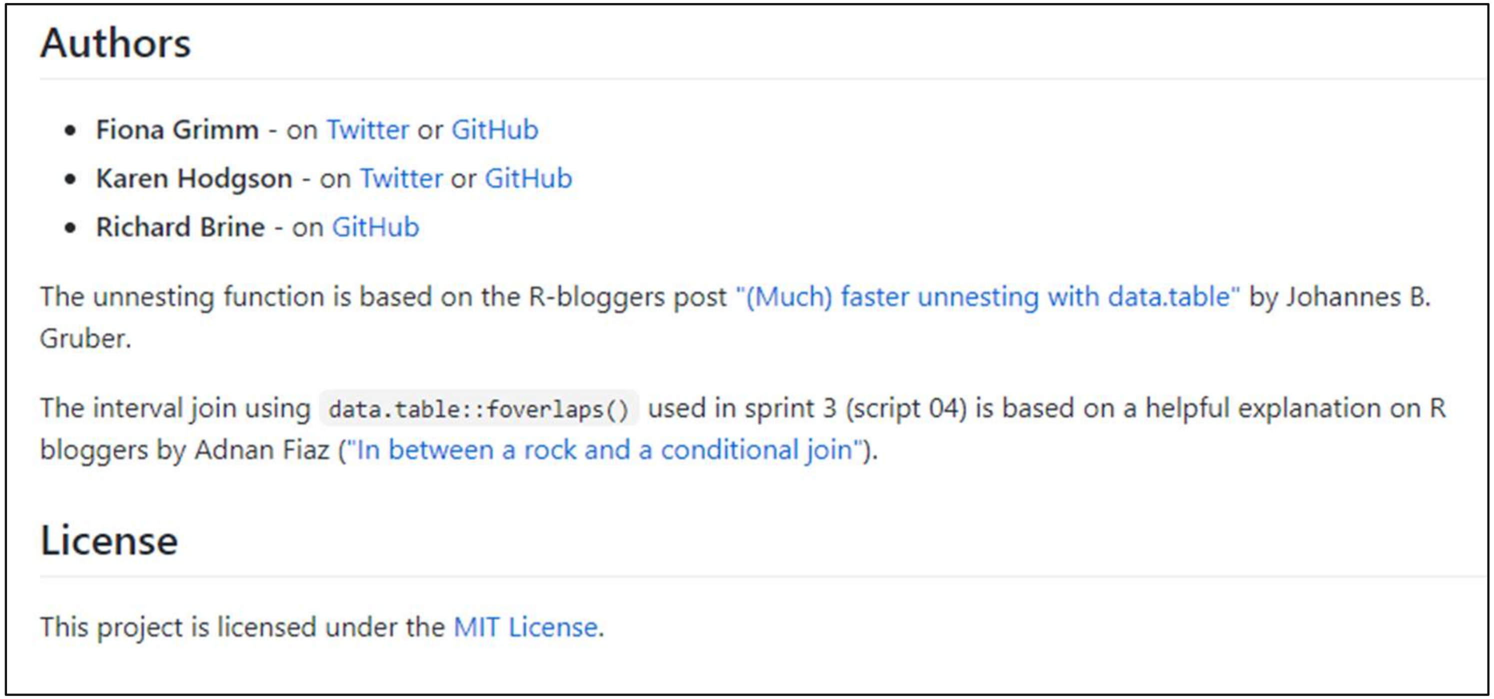
- ✓ No reason not to publish owner derived code
- ✓ Ownerships not terribly problematic
- ✓ Don’t bury away in massive user guide
- ✓ Make machine-readable – e.g. not in pdf format
- ✓ Easy to create a citation; Get a DOI?
- ✓ Make available in safe setting (e.g. project-based GitLab; dataset level Gitlab; Remote execution)

- ❖ Need to devote dedicated resources to document well
- ❖ Should become BAU

Researchers' GitHub code library

The Health Foundation's Health Analytics Lab

COVID-19 risk and health care needs of care home residents in England



Authors

- **Fiona Grimm** - on Twitter or GitHub
- **Karen Hodgson** - on Twitter or GitHub
- **Richard Brine** - on GitHub

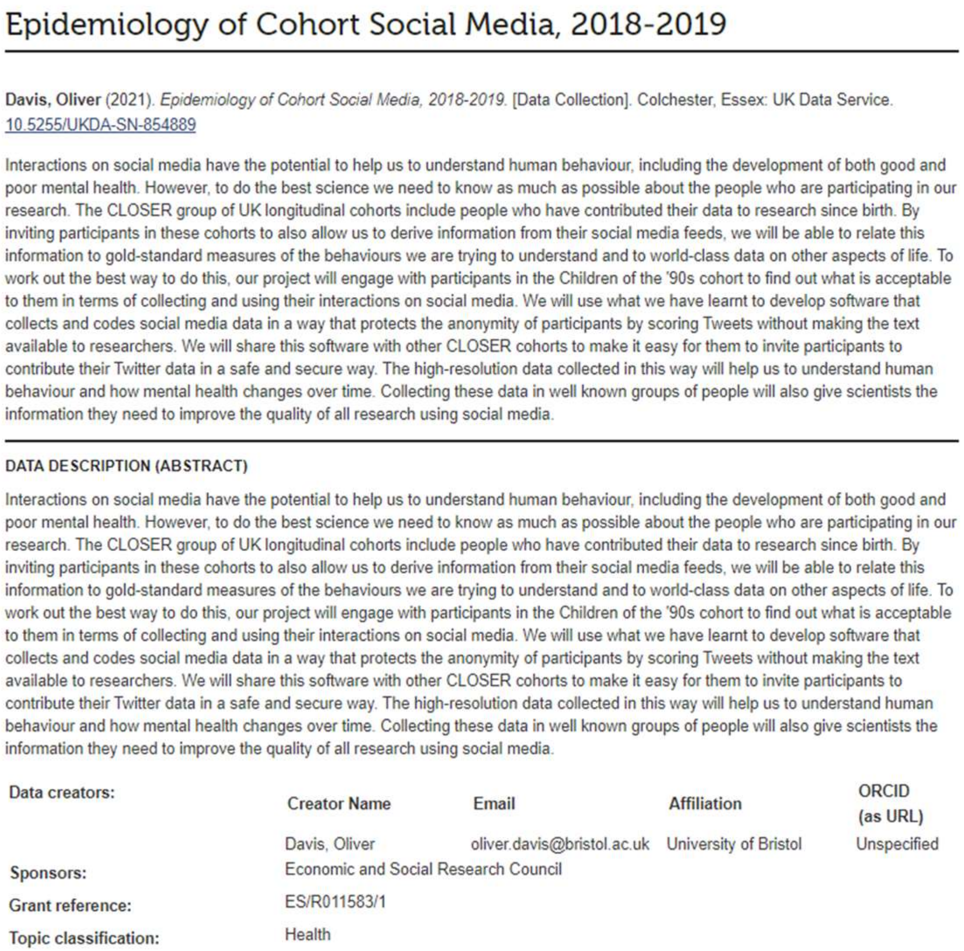
The unnesting function is based on the R-bloggers post "(Much) faster unnesting with data.table" by Johannes B. Gruber.

The interval join using `data.table::foverlaps()` used in sprint 3 (script 04) is based on a helpful explanation on R bloggers by Adnan Fiaz ("In between a rock and a conditional join").

License

This project is licensed under the MIT License.

Researcher's published code



Epidemiology of Cohort Social Media, 2018-2019

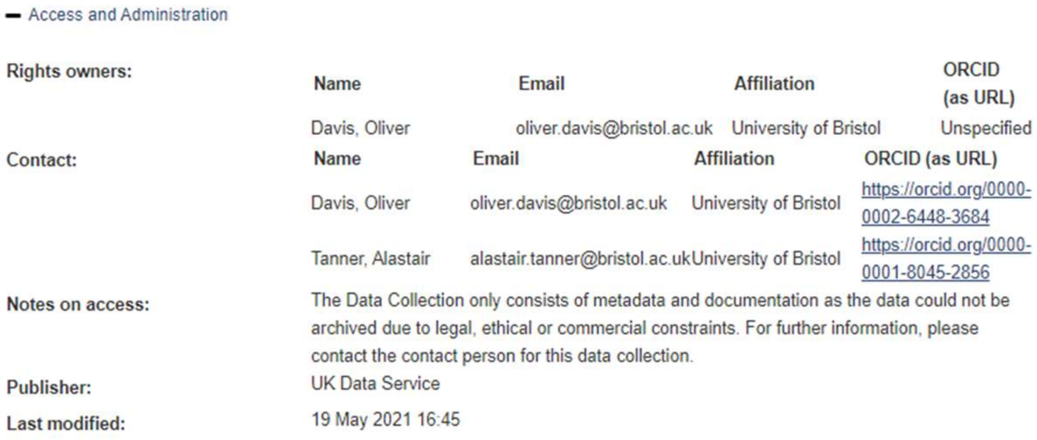
Davis, Oliver (2021). *Epidemiology of Cohort Social Media, 2018-2019.* [Data Collection]. Colchester, Essex: UK Data Service. 10.5255/UKDA-SN-854889

Interactions on social media have the potential to help us to understand human behaviour, including the development of both good and poor mental health. However, to do the best science we need to know as much as possible about the people who are participating in our research. The CLOSER group of UK longitudinal cohorts include people who have contributed their data to research since birth. By inviting participants in these cohorts to also allow us to derive information from their social media feeds, we will be able to relate this information to gold-standard measures of the behaviours we are trying to understand and to world-class data on other aspects of life. To work out the best way to do this, our project will engage with participants in the Children of the '90s cohort to find out what is acceptable to them in terms of collecting and using their interactions on social media. We will use what we have learnt to develop software that collects and codes social media data in a way that protects the anonymity of participants by scoring Tweets without making the text available to researchers. We will share this software with other CLOSER cohorts to make it easy for them to invite participants to contribute their Twitter data in a safe and secure way. The high-resolution data collected in this way will help us to understand human behaviour and how mental health changes over time. Collecting these data in well known groups of people will also give scientists the information they need to improve the quality of all research using social media.

DATA DESCRIPTION (ABSTRACT)

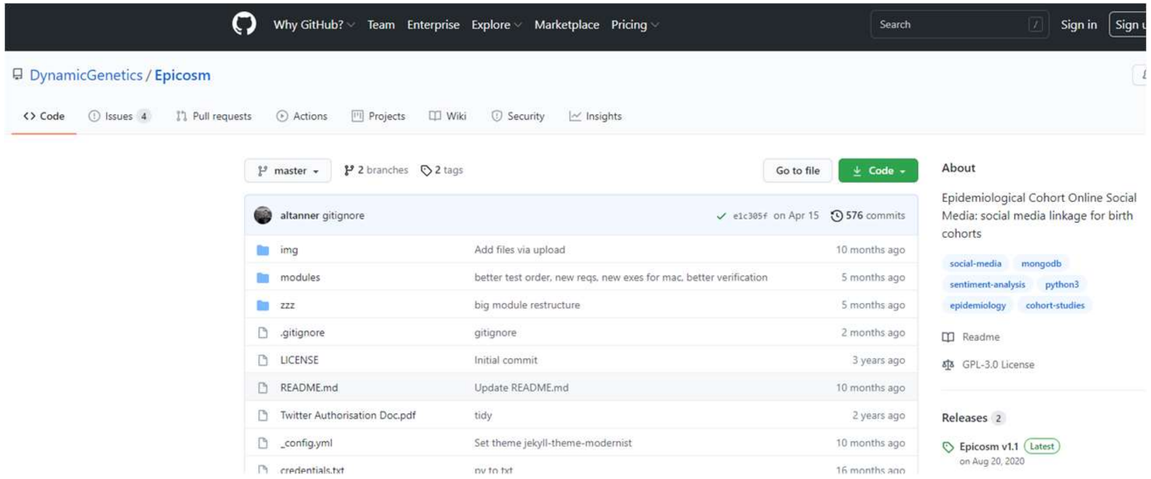
Interactions on social media have the potential to help us to understand human behaviour, including the development of both good and poor mental health. However, to do the best science we need to know as much as possible about the people who are participating in our research. The CLOSER group of UK longitudinal cohorts include people who have contributed their data to research since birth. By inviting participants in these cohorts to also allow us to derive information from their social media feeds, we will be able to relate this information to gold-standard measures of the behaviours we are trying to understand and to world-class data on other aspects of life. To work out the best way to do this, our project will engage with participants in the Children of the '90s cohort to find out what is acceptable to them in terms of collecting and using their interactions on social media. We will use what we have learnt to develop software that collects and codes social media data in a way that protects the anonymity of participants by scoring Tweets without making the text available to researchers. We will share this software with other CLOSER cohorts to make it easy for them to invite participants to contribute their Twitter data in a safe and secure way. The high-resolution data collected in this way will help us to understand human behaviour and how mental health changes over time. Collecting these data in well known groups of people will also give scientists the information they need to improve the quality of all research using social media.

Data creators:	Creator Name	Email	Affiliation	ORCID (as URL)
	Davis, Oliver	oliver.davis@bristol.ac.uk	University of Bristol	Unspecified
Sponsors:	Economic and Social Research Council			
Grant reference:	ES/R011583/1			
Topic classification:	Health			



— Access and Administration

Rights owners:	Name	Email	Affiliation	ORCID (as URL)
	Davis, Oliver	oliver.davis@bristol.ac.uk	University of Bristol	Unspecified
Contact:	Name	Email	Affiliation	ORCID (as URL)
	Davis, Oliver	oliver.davis@bristol.ac.uk	University of Bristol	https://orcid.org/0000-0002-6448-3684
	Tanner, Alastair	alastair.tanner@bristol.ac.uk	University of Bristol	https://orcid.org/0000-0001-8045-2856
Notes on access:	The Data Collection only consists of metadata and documentation as the data could not be archived due to legal, ethical or commercial constraints. For further information, please contact the contact person for this data collection.			
Publisher:	UK Data Service			
Last modified:	19 May 2021 16:45			



DynamicGenetics / Epicosm

altanner gitignore ✓ e1c305f on Apr 15 576 commits

img	Add files via upload	10 months ago
modules	better test order, new reqs, new exes for mac, better verification	5 months ago
zzz	big module restructure	5 months ago
.gitignore	gitignore	2 months ago
LICENSE	Initial commit	3 years ago
README.md	Update README.md	10 months ago
Twitter Authorisation Doc.pdf	tidy	2 years ago
_config.yml	Set theme jekyll-theme-modernist	10 months ago
credentials.txt	py to txt	16 months ago

About

Epidemiological Cohort Online Social Media: social media linkage for birth cohorts

on Aug 20, 2020

Data owners: positive solutions

- ✓ Be receptive towards value-added products create from 'outside'
- ✓ Set up a cocreation approach - added value can contribute to the data
- ✓ Share a 'house style'/template for preparing and presenting code
- ✓ Look to Reproducible Pipelines work in own organisation e.g. GSS RAP
- ✓ Set up/contribute to an external shared environment for the data e.g. a public GitHub
- ✓ Agree a standard citation style for the user derived code
- ✓ Agree a standard disclaimer – if needed - for the user derived code
- ✓ Code sharing pilot being designed at ONS SRS

The Five Reproducibles (Rs) Framework

Reproducible People

Researchers trained to write, annotate and share good code, and learn how to review other's code. Use of ORCIDs

Reproducible Projects

Research project plans, motivations and methodology should be explicit. Accountable to legal gateways and meeting the public good

Reproducible Data

Open transparent documentation; data preparation operations and derived variables published. Use of DOIs, citation and data availability statements

Reproducible Settings

Shared approaches, e.g. DEA Accredited processors

Reproducible Outputs

All research outputs should make code available to rerun analyses. Training TRE staff on code review. Code gains DOIs, citation benefits the code creators. Data owners agree code sharing frameworks for their datasets

Contact

Thanks to ONS colleagues for input into these slides

Louise Corti
Louise.Corti@ons.gov.uk

@LouiseCorti