PUBLISHING RESEARCH DATA

HOW CAN YOU DO IT, AND WHY SHOULD YOU CARE?

A WORD FROM OUR SPONSOR …

RDM@UNSW.EDU.AU

DATA PUBLICATION – WHAT IS IT?

Publishing research data is making data/primary materials produced as part of a research activity permanently available as a research output and part of the scholarly record.

This is distinct from sharing of data between researchers without a public record or information on the availability of the data.

‘Data’ here encompasses numerical/tabular data, images, audio/video recordings and transcripts of interviews, clinical trial results, source code, methodologies and workflows, lab notebooks …

… almost everything arising from a research project outside of manuscripts/drafts of research publications.

Published data is, building on Krantz & Strasser (2014, 10.12688/f1000research.3979.3):

- **available** now and into the future, with access either open/freely available or based on objective criteria
- **documented**/described so an end user can understand what the data is and how to use it
- **citable**, like a journal article or other research output

‘The underlying goals of data publication are to enable research to be reproduced and data to be reused.’

Another framework for publication of data are the FAIR principles:

- Findable: (meta)data is discoverable via search or indexing services, and has a persistent identifier
- Accessible: (meta)data is persistently retrievable via an open protocol; authentication and authorisation okay for data, or open access
- Interoperable: data uses open or common file formats, standards, vocabularies etc
- Reusable: data has documenation on its provenance and usage, and a licence

WHY YOU SHOULD

A JOURNAL MAY REQUIRE IT

- Public Library of Science (PLoS) journals require data underlying the findings of an article be ‘fully available without restriction’ at time of publication, and for manuscripts to include a ‘data availability statement’ on how to access the data – a persistent identifier for a deposit in a public repository is the preferred method.
    - They won’t accept submissions dependent on commercial or proprietary data, or where authors are not willing to publish the data.
- Nature Research journals require supporting data be made available to editors and peer reviewers, and a data availability statement in manuscripts with deposit in a public repository preferred

Alsheikh-Ali AA et al (2011, 10.1371/journal.pone.0024357) found 88% of the 50 highest impact factor journals had a data availability statement in their submission requirements, although with a variety of requirements for making data available.

You should consider the data requirements of a journal before you submit a manuscript or proposal for publication.

YOUR FUNDER MAY ENCOURAGE IT

- ARC open access policy 2017 ‘strongly encourages’ deposit of data in ‘publicly accessible repositories’
  - ARC Discovery and Linkage grant application forms have a section for describing the planned management of data including ‘access and re-use arrangements’ following the conclusion of a project
- NHMRC open access policy 2018 ‘encourages researchers to […] share research data and associated metadata arising from NHMRC supported research’

OTHER REASONS

- More citations for your articles (maybe)
  - Colavizza et al (2019, https://arxiv.org/abs/1907.02565)'s study of ~530k Public Library of Science and BioMed Central open access articles found articles linking to a data repository record had an average 25% higher citation impact
  - Piwowar, Day and Fridsma (2007, 10.1371/journal.pone.0000308)'s study of 85 cancer microarray clinical trial publications found the 48% with publicly available data enjoyed 85% of the aggregated citations

The Australian Code for the Responsible Conduct of Research (2018) and the UNSW Research Code of Conduct require researchers to retain accurate and secure records of research data/primary materials and ‘allow access and reference to these by interested parties’.

It’s good for research:

- increases transparency by faciliating independent verification and critique of research findings
- allows for new analysis and research, including from expensive or unrepeatable data collection efforts
- discoverable and available data may inspire novel applications e.g. combining data from different disciplines

WHY YOU SHOULDN’T OR CAN’T

- publishing data precluded by how it was collected, funded etc
- lack of suitable technical infrastructure/system
- giving up exclusive access to a research asset
- other reasons???

HOW TO DO IT

DON’T DO THIS

1. Make your data downloadable via somerandomserver.unsw.edu.au, or a folder in Cloudstor/OneDrive/Dropbox/etc
2. Put the URL somewhere in the body of your journal article
3. Walk away

What you need:

- a publicly visible record for the data somewhere end users are able to find it
  - data repository, Research Data Australia, Google Dataset Search …
- metadata and documentation describing the data
- some long-term storage/hosting for the data
- some means of providing the data to end users
  - click to download, mediated access (approval workflow/password) …
- persistent identifier (e.g. DOI) or URL
- a licence for the data

Options include:

- disciplinary data repository — see re3data.org or Scientific Data’s recommended data repositories
- general data repository — Figshare, Dryad
- UNSW supported system
- roll your own solution

DATA JOURNALS

Data journals (e.g. Nature’s Scientific Data) publish articles describing publicly available datasets without any analysis or conclusions.

These are generally not considered ‘prior publication’ if the same authors seek to publish an analysis of the data.

THIS SLIDE INTENTIONALLY LEFT BLANK

PUBLISHING SENSITIVE DATA

Sensitive data is sometimes not considered for publication because of actual or perceived issues or risk with disclosing information that could compromise the privacy or safety/security of subjects (or some other bad stuff).

Sensitive data about humans commonly refers to information that can identify an individual (e.g. name, date of birth, email address) accompanied by information that is considered sensitive e.g. medical or health status/history, racial or ethnic origin, religious or political affiliation, criminal record et al.

Data not about humans can also be sensitive, e.g. location of an endangered and/or commercially valuable species of animal or plant.

Under the Australian Privacy Act 1988, sensitive data containing personal identifying information generally cannot be shared. However if the data is deidentified or measures applied such that individuals are not ‘reasonably identifiable’, this protection does not apply.

Data can be intrinsically sensitive, or sensitive by context and/or combination with other data.

Examples include determining an identity or some sensitive information through **triangulation** of multiple non-sensitive pieces of information, or **data linkage** of two or more datasets containing information about the same individual.

Statistical disclosure control (SDC) methods involve modifying data (both ‘microdata’ about individuals and aggregated) to reduce potential for disclosure of identities or sensitive information, including perturbation of data or omitting data points assessed to be sensitive.

For new research projects, you should obtain consent for data publication from participants – explain how, where and when you intend to publish/make the data available, and how you will deidentify or manage the risk of disclosures from the data.

You should also secure approval from the relevant UNSW ethics committee.

For existing data/completed projects, you may be able to publish if:

- consent from participants doesn’t preclude publishing/making the data available
- obtaining consent for data publication isn’t practical
- data has been deidentified consistent with the Privacy Act 1988

(This section mostly from the ANDS guide to Publishing and sharing sensitive data. Go read this, it’s really good.)

COPYRIGHT AND LICENSING

Data that is published/made available without a licence cannot (legally) be reused.

The utility of published data is maximised by applying an open content licence that allows the data to be used and redistributed with no or minimal restrictions.

A popular option for research data is the Creative Commons suite of open content licences.

The base ‘Attribution’ licence allows the data to be used for any purpose, and to be redistributed, provided this usage is attributed.

Variant licences apply additional conditions e.g. prohibiting use for commercial purposes.

You may encounter suggestions that ‘copyright doesn’t apply to data’, or that all research data should be put in the public domain.

You should seek advice before putting any UNSW data into the public domain.

UNSW SERVICES AND SUPPORT

ResData offers functionality to create a dataset metadata record in Research Data Australia, and optionally mint a DOI. You can upload data files and make them publicly accessible (no mediated access), with limits of 2G per file and 20G total per dataset submission.

rdm@unsw.edu.au can provide advice and recommendations on all aspects of data publication. The outreach librarian for your school or research centre can advise on discipline-specific options and conventions for publishing data.

FIN

PLEASE BOMBARD PRESENTER WITH QUESTIONS