Digital Tools for Reproducible Research

Materials https://bolibaugh.github.io/DigitalTools/ https://osf.io/jrxyw/

Contact cylcia.bolibaugh@york.ac.uk | @CBolibaugh

Research increasingly reliant on computational and data skills…

“Some other time”

Too many options

Inefficient to learn on your own

Shaming?

Thanks to Rochelle Terman for the motivation.

Research increasingly reliant on computational and data skills...

"Some other time"

Reproducible workflows (including code) save time once set up

"Too many options"

Learn the logic now, specialise later

Inefficient to learn on your own

True -- do it here

Shaming? Everyone started somewhere; very few are experts

Week 5 Reproducible research 15 May 2019

Week 6 Preregistration 22 May 2019

Week 8 Open data 5 June 2019

Week 9 Reproducible analyses, power analysis and simulation 12 June 2019

Week 10 Writing a reproducible manuscript 19 June 2019

Thanks to Rochelle Terman for the motivation.

03 Open Data

Today

Ethics, legality & FAIR principles in Open Data

Manage your data so that it can be archived and shared to the fullest extent permitted by legal & ethical frameworks.

@CBolibaugh

Tasks

1. Learn about FAIR data principles
2. Ensure data is human & machine readable
3. Create meta-data
4. Identify relevant repositories
5. Assign a DOI & license
6. Learn about ethical & legal issues
7. Browse example datasets

By the end of the today, you should be able to structure your data for sharing, identify an appropriate repository & license, & understand relevant legal & ethical frameworks.

Thanks to The Turing Way Guide to Reproducible Data Science.

Why share data?

- validate published research findings
- enable data re-use, and the combination of datasets from multiple sources
- receive credit for research outputs other than publications
- comply with University RDM Policy

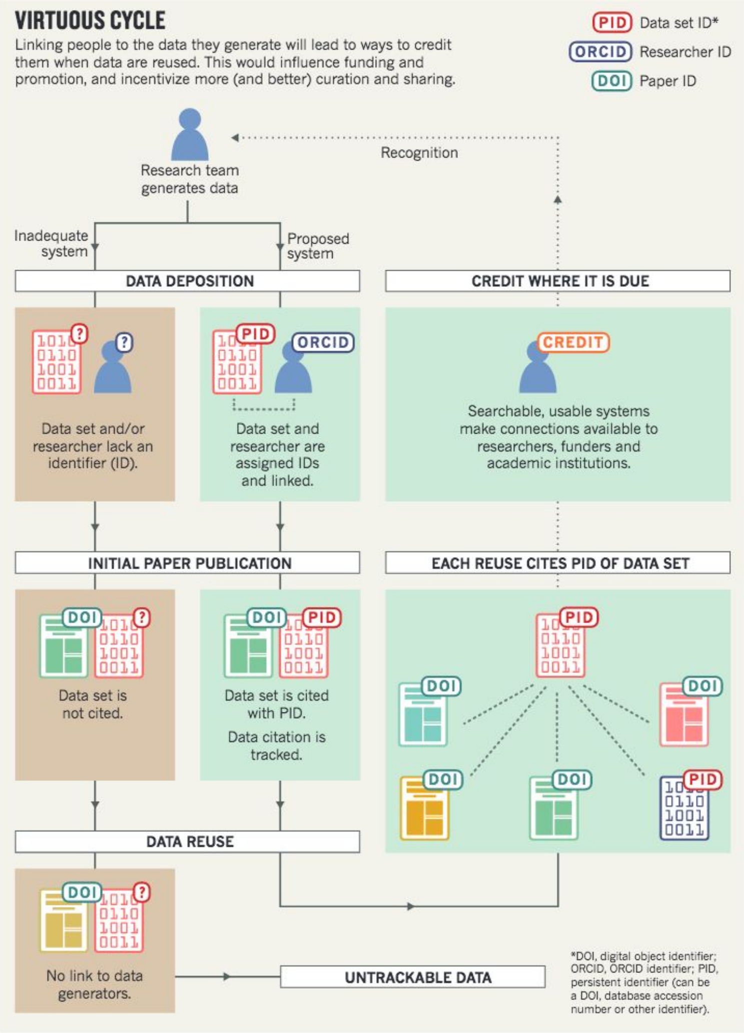
Publicly funded research data are a public good, produced in the public interest, which should be made openly available with as few restrictions as possible in a timely and responsible manner.

UKRI Research Council common principles on data, Principle 1

@CBolibaugh

Why share data?

- ➔ validate published research findings
- ➔ enable data re-use, and the combination of datasets from multiple sources
- ➔ receive credit for research outputs other than publications



VIRTUOUS CYCLE

Linking people to the data they generate will lead to ways to credit them when data are reused. This would influence funding and promotion, and incentivize more (and better) curation and sharing.

PID Data set ID*
ORCID Researcher ID
DOI Paper ID
Research team generates data
Recognition
Inadequate system
Proposed system
DATA DEPOSITION
CREDIT WHERE IT IS DUE
Data set and/or researcher lack an identifier (ID).
Data set and researcher are assigned IDs and linked.
CREDIT
Searchable, usable systems make connections available to researchers, funders and academic institutions.
INITIAL PAPER PUBLICATION
EACH REUSE CITES PID OF DATA SET
Data set is not cited.
Data set is cited with PID.
Data citation is tracked.
DATA REUSE
No link to data generators.
UNTRACKABLE DATA

*DOI, digital object identifier; ORCID, ORCID identifier; PID, persistent identifier (can be a DOI, database accession number or other identifier).

FAIR data principles

The FAIR guiding principles for scientific data management and stewardship (Wilkinson et al., 2016):

- ➔ **Findable:** the first step in (re)using data is to find them, and descriptive metadata is essential.
- ➔ **Accessible:** are data open with no restrictions, or is authentication and authorisation necessary?
- ➔ **Interoperable:** data need to be integrated and/or interoperable with existing standards
- ➔ **Reusable:** data should be well-described so that they can be used or replicated in different settings

Making data ‘FAIR’ is not the same as making it ‘open’ (as accessibility principle explains).

Data should be as open as possible, as closed as necessary.

FAIR data principles

So, how can I make my research data FAIR?

1. Ensure your data is in a simple, standard format or formats which is machine and human readable.
2. Check, reformat or create metadata to clearly describe what the data is, how it was collected, and any associated strengths/weaknesses to someone that finds it.
3. Identify a relevant, easily discoverable repository or repositories to host your data, and upload it there.
4. Assign your data a persistent identifier such as a DOI, & appropriate license.

Plan for archiving and sharing your data from the beginning of your project, so you can ensure you comply with ethical and legal frameworks.

Human- and machine-readable data

(or how to avoid the dangers of spreadsheets)

For tabular data:

Make sure all raw data is read-only

Follow principles of good organisation for any spreadsheet type data

Save data in plain text files (.csv, or .tsv), rather than excel, or spss

Don't do these things

For other data, make sure to follow any community standards like CHILDES

Broman, K. W., & Woo, K. H. (2018). Data organization in spreadsheets. *The American Statistician*, *72*(1), 2–10. https://doi.org/10.1080/00031305.2017.1375989

- Be consistent
- Write dates as YYYY-MM-DD
- Don't leave any cells empty
- Put just one thing in a cell
- Organize the data as a single rectangle
- Don't include calculations in the raw data files
- Don't use font color or highlighting as data
- Choose good names for things
- Make backups
- Use data validation to avoid data entry mistakes

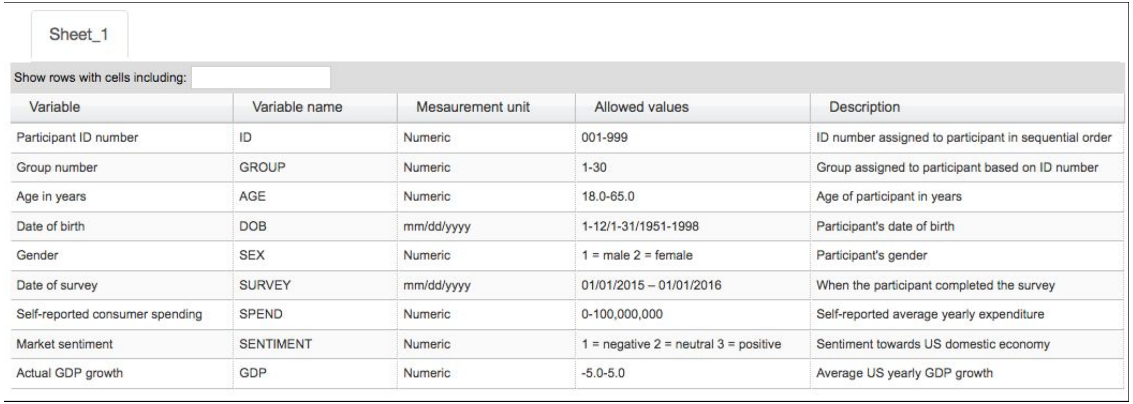
@CBolibaugh

Creating meta-data (1)

Having data available is of no use if it cannot be understood. Check, reformat or create metadata to clearly describe what the data is, how it was collected

- Variables should be defined and explained using [data dictionaries](#)

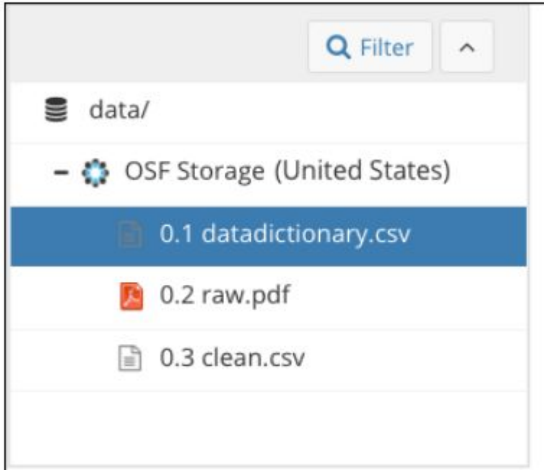
- Data should be stored in logical and hierarchical folder structures with a README file used to describe the structure.



Sheet_1

Show rows with cells including:

Variable	Variable name	Mesaurement unit	Allowed values	Description
Participant ID number	ID	Numeric	001-999	ID number assigned to participant in sequential order
Group number	GROUP	Numeric	1-30	Group assigned to participant based on ID number
Age in years	AGE	Numeric	18.0-65.0	Age of participant in years
Date of birth	DOB	mm/dd/yyyy	1-12/1-31/1951-1998	Participant's date of birth
Gender	SEX	Numeric	1 = male 2 = female	Participant's gender
Date of survey	SURVEY	mm/dd/yyyy	01/01/2015 – 01/01/2016	When the participant completed the survey
Self-reported consumer spending	SPEND	Numeric	0-100,000,000	Self-reported average yearly expenditure
Market sentiment	SENTIMENT	Numeric	1 = negative 2 = neutral 3 = positive	Sentiment towards US domestic economy
Actual GDP growth	GDP	Numeric	-5.0-5.0	Average US yearly GDP growth



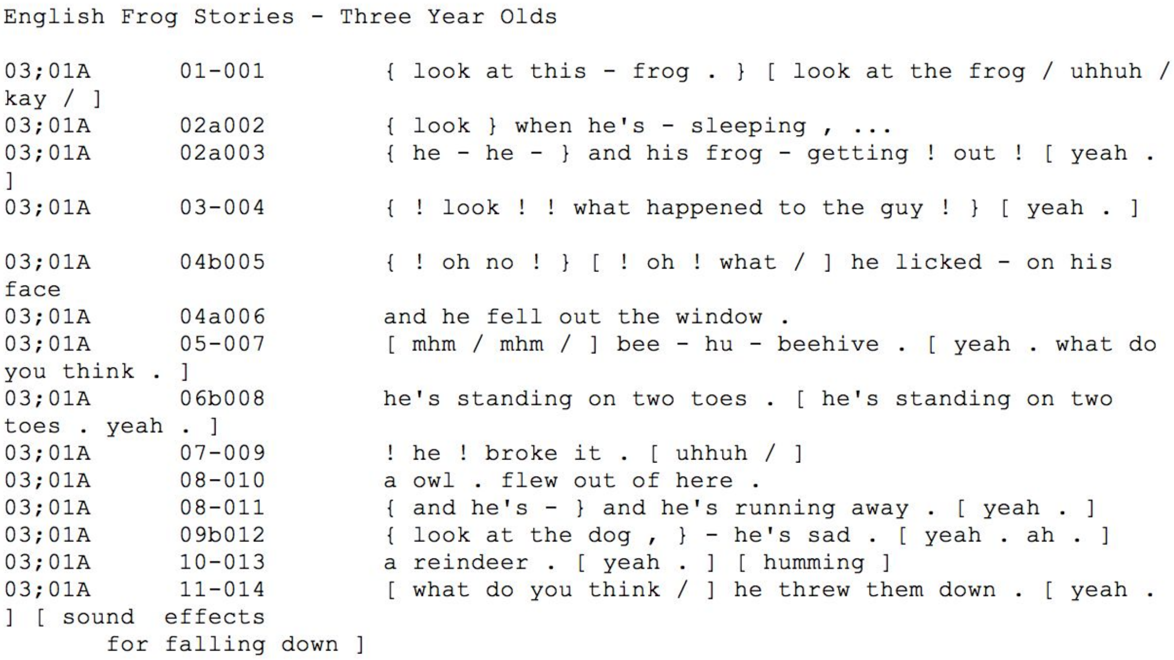
Creating meta-data (2)

Use recognised community (meta)data standards to make it easier for datasets to be combined.

For example, for brain data the Brain Imaging Data Structure is the standard to use, and CHILDES/CHAT conventions have been widely adopted for language production data.

There are currently no standards for L2 data, though a COST action is being planned to develop these. IRIS ontology maybe a useful starting point.

For eyetracking data, a BIDS extension is being developed.



```
English Frog Stories - Three Year Olds

03;01A      01-001       { look at this - frog . } [ look at the frog / uhhuh /
kay / ]
03;01A      02a002       { look } when he's - sleeping , ...
03;01A      02a003       { he - he - } and his frog - getting ! out ! [ yeah .
]
03;01A      03-004       { ! look ! ! what happened to the guy ! } [ yeah . ]

03;01A      04b005       { ! oh no ! } [ ! oh ! what / ] he licked - on his
face
03;01A      04a006       and he fell out the window .
03;01A      05-007       [ mhm / mhm / ] bee - hu - beehive . [ yeah . what do
you think . ]
03;01A      06b008       he's standing on two toes . [ he's standing on two
toes . yeah . ]
03;01A      07-009       ! he ! broke it . [ uhhuh / ]
03;01A      08-010       a owl . flew out of here .
03;01A      08-011       { and he's - } and he's running away . [ yeah . ]
03;01A      09b012       { look at the dog , } - he's sad . [ yeah . ah . ]
03;01A      10-013       a reindeer . [ yeah . ] [ humming ]
03;01A      11-014       [ what do you think / ] he threw them down . [ yeah .
] [ sound  effects
       for falling down ]
```

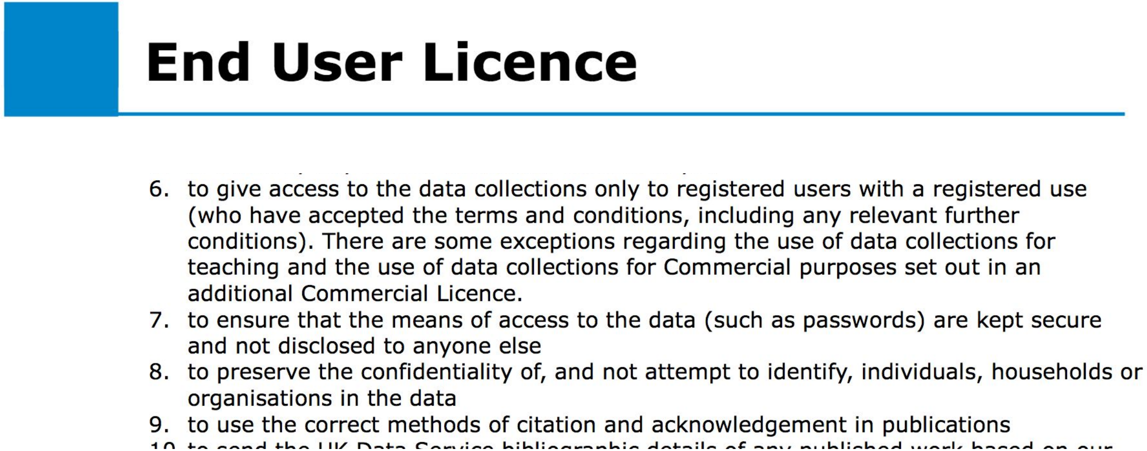
From R. A. Berman & D. I. Slobin (1994). Relating events in narrative: A crosslinguistic developmental study. Hillsdale, NJ: Lawrence Erlbaum Associates.

Identifying a repository

UK Data Service

Registry of Research Data Repositories: https://www.re3data.org/

- OSF for data that can be shared without restriction (or kept private)
- IRIS for L2 data that can be shared without restriction
- UKDS (e.g. ReShare, or QualiBank) for access controls via EULs
- York Research Database if no alternative; also has access control:



End User Licence

6. to give access to the data collections only to registered users with a registered use (who have accepted the terms and conditions, including any relevant further conditions). There are some exceptions regarding the use of data collections for teaching and the use of data collections for Commercial purposes set out in an additional Commercial Licence.
7. to ensure that the means of access to the data (such as passwords) are kept secure and not disclosed to anyone else
8. to preserve the confidentiality of, and not attempt to identify, individuals, households or organisations in the data
9. to use the correct methods of citation and acknowledgement in publications
10. to send the UK Data Service bibliographic details of any published work based on our

*

Persistent identifiers (DOI, and license)

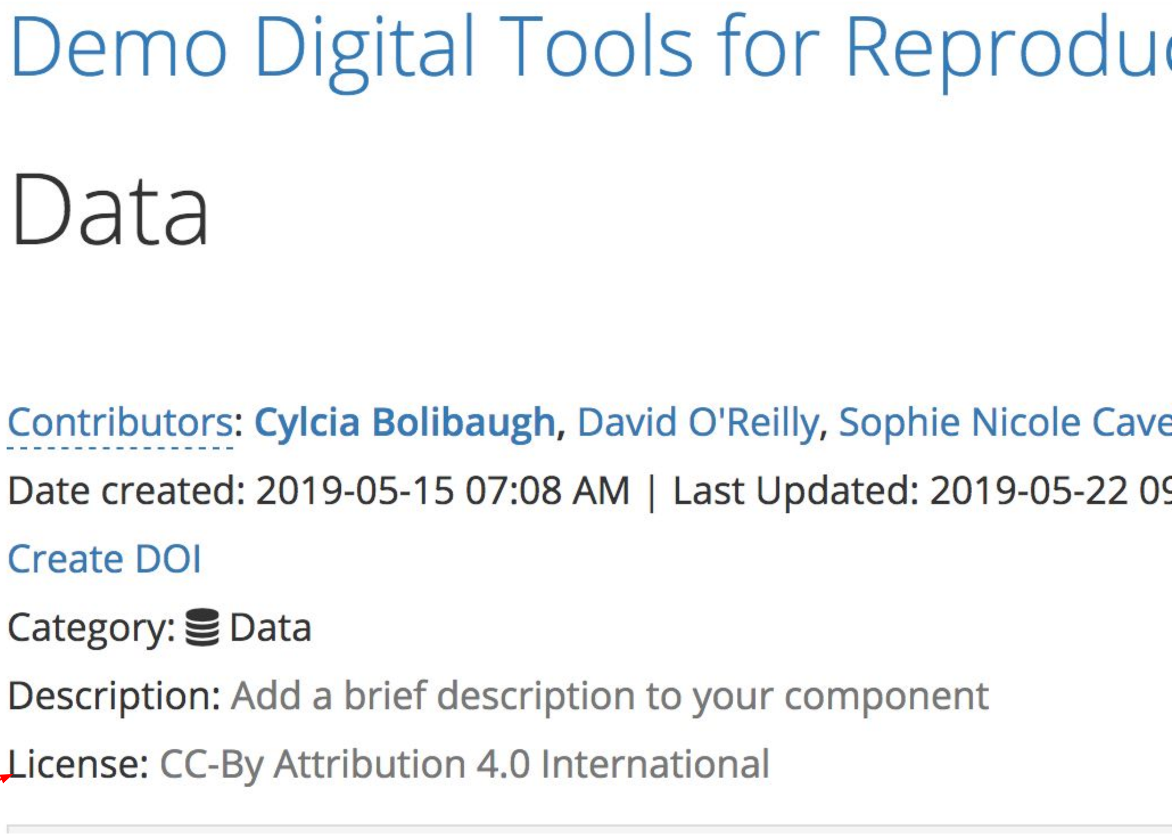
If uploading to OSF, select ‘Create DOI’.

If uploading to IRIS, register in PURE and request DOI from library support.

If uploading to UKDS, DOI will be assigned

Don’t request a DOI until you have a dataset you are happy to share.

If dataset is open, consider adding a CC-BY license for reuse with credit, or learn more about licenses for data.



Demo Digital Tools for Reprodu

Data

Contributors: **Cylcia Bolibaugh**, David O'Reilly, Sophie Nicole Cave

Date created: 2019-05-15 07:08 AM | Last Updated: 2019-05-22 09

Create DOI

Category: Data

Description: Add a brief description to your component

License: CC-By Attribution 4.0 International

Legal & ethical issues

Planning for data sharing from the beginning of your project:

- UKDS: Legal and ethical issues

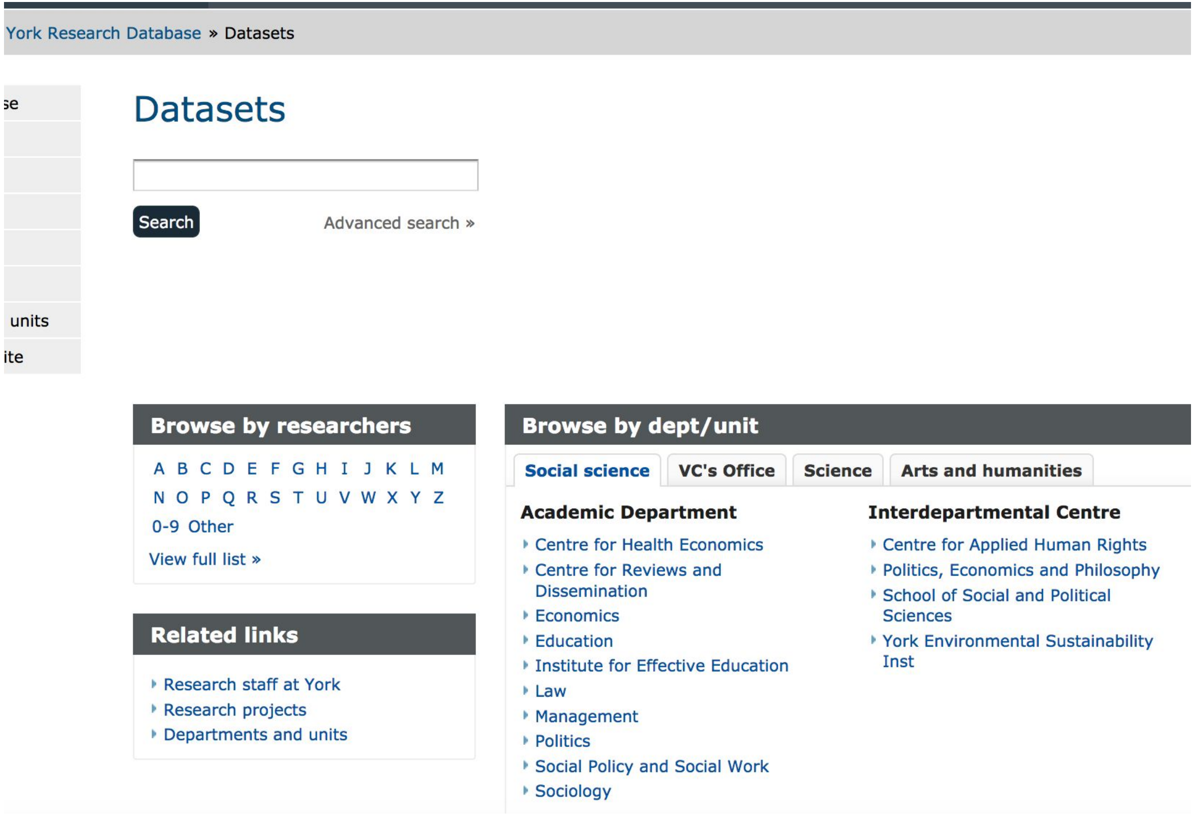
  Advice from the UK Data Service about managing research data about people, including informed consent, anonymisation and access control.

- University RDM web pages: Ethical and legal issues
  Guidance on the management of confidential, sensitive and/or personal data; includes links to University guidance on Data Protection, Freedom of Information and Intellectual Property Rights.

- UKRI: Guidance on best practice in the management of research data
  UK Research and Innovation (formerly RCUK) recognises that there are legal, ethical and commercial constraints on release of research data. This document provides useful guidance on UKRI's expectations in relation to legal, ethical and commercial constraints.

Find datasets

- Browse by department: York Research Database
- IRIS for L2 research; search everything and use ‘data’ as type of material
- Explore Qualibank



York Research Database » Datasets

se

units

ite

Datasets

Search Advanced search »

Browse by researchers

A B C D E F G H I J K L M
N O P Q R S T U V W X Y Z
0-9 Other

View full list »

Related links

- Research staff at York
- Research projects
- Departments and units

Browse by dept/unit

Social science | VC's Office | Science | Arts and humanities

Academic Department

- Centre for Health Economics
- Centre for Reviews and Dissemination
- Economics
- Education
- Institute for Effective Education
- Law
- Management
- Politics
- Social Policy and Social Work
- Sociology

Interdepartmental Centre

- Centre for Applied Human Rights
- Politics, Economics and Philosophy
- School of Social and Political Sciences
- York Environmental Sustainability Inst

Explore two datasets

Compare two openly available datasets:

- Example project from Eva Poort & Jenni Rodd:
    - The project page https://osf.io/ndb7p/
    - Task one: how is it licensed?
    - Task two: Where is the metadata?
- Example project from Nagle et al (Nagle, C., Trofimovich, P., & Bergeron, A. (in press, 2019). Toward a dynamic view of second language comprehensibility. Studies in Second Language Acquisition):
    - The project page: https://osf.io/97kur/
    - Task one: how is it licensed?
    - Task two: Where is the metadata?
    - Task three: Could the demographic data be openly posted under the terms of the UK Data Protection Act, and GDPR?

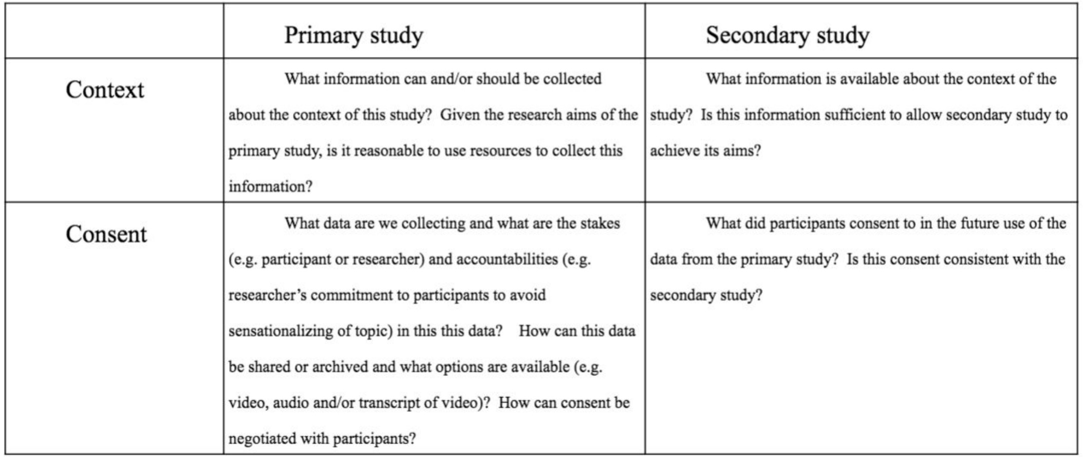
Resources

Qualitative data

Chauvette, A., Schick-Makaroff, K., & Molzahn, A. E. (2019). Open Data in Qualitative Research. *International Journal of Qualitative Methods*, *18*, 1609406918823863.

https://journals.sagepub.com/doi/full/10.1177/1609406918823863

Branney, P., Reid, K., Frost, N., Coan, S., Mathieson, A., & Woolhouse, M. (2018, October 31). A meta-framework for designing open data studies in psychology: ethical and practical issues of open qualitative data sets. https://doi.org/10.1080/14780887.2019.1605477



	Primary study	Secondary study
Context	What information can and/or should be collected about the context of this study? Given the research aims of the primary study, is it reasonable to use resources to collect this information?	What information is available about the context of the study? Is this information sufficient to allow secondary study to achieve its aims?
Consent	What data are we collecting and what are the stakes (e.g. participant or researcher) and accountabilities (e.g. researcher's commitment to participants to avoid sensationalizing of topic) in this this data? How can this data be shared or archived and what options are available (e.g. video, audio and/or transcript of video)? How can consent be negotiated with participants?	What did participants consent to in the future use of the data from the primary study? Is this consent consistent with the secondary study?

Figure 2. Context and consent meta-framework for open data

You can look through RAD presentation from Maureen Haaker from UKDS re sharing qualitative data here.

Resources

Quantitative data

Amnesia data anonymization tool

Meyer, M. N. (2018). Practical tips for ethical data sharing. *Advances in Methods and Practices in Psychological Science*, *1*(1), 131-144.https://doi.org/10.1177/2515245917747656

You can look through previous EROS presentations from Open Data events here.

Follow-up

Preparation for Week 9

Select either: (1) one of your own (completed or planned) empirical research studies, or (2) an empirical paper that interests you.

Familiarise yourself sufficiently with the key variables in your study so that you can simulate data for your in the next session.