Publishing and citing research data

Louise Corti
UK Data Archive
University of Essex

2015 Big Data and Analytics Summer School
Colchester
24 August, 2015
Overview

While research data is often exchanged in informal ways with collaborators and colleagues, formally publishing data brings many advantages. Data publishing has grown rapidly in recent years, and there are myriad places to do so. A significant advantage to formal publication is enabling formal citation and proper attribution.

Areas of coverage

- Why share data?
- Where to publish data
- Types of repository
- Data journals
- Citing data
- Persistent identifiers
Benefits of data sharing

To funders
• make optimal use of publicly funded research
• avoid duplication of data collection
• maximise return on investment

To the scholarly community
• maintain professional standards of open inquiry
• maximise transparency where appropriate
• quality improvement from verification, replication and trust
• documentation for research design and teaching
• promote innovation through unintended, new uses of data
• develop long time series of data
Benefits of data sharing

To research participants
• allow maximum use of their contributed data/information
• minimise data collection on the hard-to-reach (e.g. ill, elites)

To the public
• production of high quality research with social value
• advance science to the benefit of society
• compliance with laws and regulations
• adoption of emerging norms – ‘open access’ publishing
• to be, and appear to be, open and accountable
International Funder policies

• Largely based on the OECD Principles and Guidelines for Access to Research Data from Public Funding

• UK - variety of models
  • Data management plans and recommendation only
  • Dedicated data centres
  • Institutions taking responsibility

• Europe (European Research Council/Horizon 2020)
  • Pilot on open access to research data
  • Data management guidelines for Horizon 2020 (~ policies)
  • Communication & recommendation on access to / preservation of scientific information (publications, data)

• USA data management plans, e.g NSF and NIH
Journal / Publisher Data Policies

• Many science journals (Science, Nature, PLOS ONE etc.) have data policies relating to data sharing
  “PLOS ONE will not consider a study if the conclusions depend solely on the analysis of proprietary data” … “the paper must include an analysis of public data that validates the conclusions so others can reproduce the analysis.”
• e.g. BioMed Central open data statement

• Data underpinning publication accessible
  • upon request from author
  • as supplement with publication
  • in public repository
  • in mandated repository (e.g. PANGAEA – Elsevier)

• Citation via unique DOIs
• JORD project: survey of journal policies
Data sharing – a shared responsibility

- Funders: provide policies, mandates and some infrastructure funding
- Funded researchers: create, manage and use data
- Departments/centres: provide local support and some infrastructure
- Institutions: provide a supporting framework
  - grant-application and funding support
  - research integrity framework
  - IT and data storage facilities
  - Data management guidance and training
- Clarify roles and responsibilities early on
NSF & NIH Funded Data Collections: Where are they today? N=1,544

- Data Are Archived: 14.2%
- Has Copy of Data: 58.7%
- Data Are Lost: 25.7%

Source: Pienta, Gutmann, & Lyle, 2009
Why are data not shared?

- Preparing data and documentation can be enormously time consuming
- Limited resources for data preparation
- Need to protect the confidentiality of respondents
- Fear of getting “scooped”
- Lack of rewards for sharing

Source: Pienta, Gutman, & Lyle, 2009
Data sharing pros and cons

- Pros and cons

- Some reasons: Reasonsnottoshare_exercise.pdf
Where can I share my data?

- Ad hoc
- Web site
- Institutional repository
- Domain Data Centre
- Journal extra material
- Cloud solutions
Trustworthy institutionalized venues

- **Usable**: Include content that is in well-established machine-actionable digital formats
- **Discoverable**: Provide a publicly available and freely searchable catalogue
- **Meaningful**: Offer metadata, documentation and other materials needed to facilitate data re-use
- **Citable**: Facilitate formal data citation by assigning persistent, globally resolvable, machine-actionable identifiers that verifiably link to specific versions of data
- **Secure**: Maintain policies and procedural controls to protect confidentiality and personal privacy as required by law and the ethical standards of the research community
- **Durable**: Ensure long-term preservation of and access to their digital assets
Discipline specific repositories and data centres

- Dedicated to archiving, preserving and disseminating discipline specific digital data
  e.g. UK Data Archive, Environmental Information Data Centre, Visual Arts Data Service, Endangered Language Archive

- UK Data Service has many components:
  - Main collection is actively curated but reserved for high impact data, primarily from government departments
  - UK Data Service ReShare is the option for most research data offered to UK Data Service
Data journals

- A fairly new phenomenon, but growing
- Publish a detailed journal style article describing the data and how it was collected
- Recommends or provides a place of deposit e.g. Nature Scientific Data (http://www.nature.com/scientificdata/)
Domain repository: UK Data Service ReShare

- [http://reshare.ukdataservice.ac.uk/](http://reshare.ukdataservice.ac.uk/)
- A platform through which to get your research data published
- It’s remit is primarily data produced by UK social science researchers, but also other social science research data of interest
- Variety for license from open to closed or embargoed
Easy to publish and upload data

Edit collection: Data Collection #851515

To deposit a data collection, you must accept the ReShare Terms and Conditions.

- I confirm that I am the owner of the copyright and associated intellectual property rights in the whole Data Collection or am otherwise lawfully entitled to grant this licence on behalf of each and every owner;
- I grant a non-exclusive, royalty-free licence to the UK Data Archive (a department of the University of Essex and not a separate legal entity) of Wivenhoe Park, Colchester, CO4 3SQ (the “University”) to hold, make copies of, and disseminate copies of the Data collection, in accordance with the access conditions I will specify when uploading data files; open data accessible to users without registration; or safeguarded data accessible to users registered with the data services provided by the UK Data Archive.

I agree to the ReShare data deposit terms and conditions

Save for later  Cancel  Next >
Common issues in self-publishing data

Data
• Messy file names
• Are the file names disclosive?
• Is there any disclosive information in the file properties?
• Inconsistent layout of bio info in data files
• Use of unagreed track changes in text documents
• Formats – suitable format?

Documentation
• Messy file names
• Published articles cannot be shared, but instead simply referenced
• Watch out for dead links to related articles or reports!
Citing data

- Citation a fundamental part of research and academia in general
- Just as articles are cited, data which has contributed to research should be cited

- Data citation:
  - Fairly acknowledges the authors sources
  - Promotes reproduction of research findings
  - Makes it easier to find data for others who are interested
  - Allows impact of data to be tracked
  - Provides a structure that recognises and rewards data creators
Data citation conventions

- Accurately citing data is dependent on high quality metadata
- Historically, data citation has tended to take form of an imprecise acknowledgement,
- Now publishers are adopting more formal conventions

- No one standard citation format has emerged
- But certain things can be identified as crucial for data in particular:
  - the precise version of the dataset used e.g. edition number or release date info
  - where to find the data i.e. a URL
  - or even better, a persistent identifier (DOI)
UK Data Service citation convention

• UK Data Service specifies that “All works which use or refer to these materials [the data collection] should acknowledge these sources by means of bibliographic citation”

• Our format is:

• The last part of the above is a special type of URL called a Digital Object Identifier (DOI), a type of persistent identifier
• Clicking on this enables a third party look up of the DOI and points you to its current location
• Regardless of the data collection current online location, this URL will always allow you to find it
Understanding Society: Waves 1-2, 2009-2011

Persistent identifier: 10.5255/UKDA-SN-6614-4
Connecting research and researchers

- Research in the digital realm is becoming increasingly linked up
- Leverage this to increase your profile

- Get an ORCID (Open Researcher and Contributor ID) and identify yourself as a unique researcher
- ORCID provides a persistent digital identifier that distinguishes you from every other researcher i.e. *that* Dr. John Smith
- Looks something like: http://orcid.org/xxxx-xxxx-xxxx-xxxx
- Simple and free to register at: http://orcid.org/
Link together your research

Source: ORCID: Connecting Research and Researchers,
Biblioteca del Campus Terrassa on Jul 11, 2013
Peer review of data

• Increasing in popularity
• Journals doing this - replicability agenda
• No one single standard for ‘quality’
• Make metadata quality explicit:
  • Collection description
  • Data description: file and variable names & labels
  • Relationships between tables/files
  • Provenance of data and methods
What is research transparency

• Data Access and Research Transparency (DA-RT): A Joint Statement by Political Science Journal Editors

  “Journal editors commit their respective journals to the principles of data access and research transparency, and to implementing policies requiring authors to make as accessible as possible the empirical foundation and logic of inquiry of evidence-based research”


• Few established rules yet for implementation
Transparency Organizations
Justification for and Benefits of Transparency

- Compliance with a methodological approach is compelled by an epistemological need
  - social science is process-dependent
- And rule-following must be public
- The process is part of the outcome
- demonstrating the processes that produced a conclusion enables and empowers evaluation
- social science is process-visibility dependent
# Process Dependence vs. Process-Visibility Dependence

<table>
<thead>
<tr>
<th></th>
<th>Process Dependent</th>
<th>Process-Visibility Dependent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cuisine</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Movie CGI for Audience</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Movie CGI for Master Class</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Social Science</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Data are a Public Good and a Prerequisite for Transparency

Data made available for secondary analysis

Data used to support an evidence-based claim
Quantitative data transparency

- Exists for journals in psychology, economics and political science
- Submit data and syntax
- Journal’s own repository, e.g. dataverse
  https://dataverse.harvard.edu/dataverse/ajps
The American Journal of Political Science is committed to significant advances in knowledge and understanding of citizenship, governance, and politics, and to the public value of political science research.
Production Transparency: Describing Data Generation

- Engagement with Social World
- Information Selection
- Information Extraction
- Data
- Reduction/Transformation
- Information Consideration
Production Transparency: Describing Data Generation

- Lab-book approach
- Carefully cataloguing and logging information-collection and data-generation processes
- Notes become “raw material” for a more succinct description
- Retrospective narrative or interview (less detail)
Which Data Need to be Shared to Achieve Transparency?

• Fewer than all, more than none (helpful…)
• Just those that were cited?
• The data that underlie the central claims? Or that underlie potentially controversial conclusions?
• Where should those data sit
• Depends on
  • Form of the data – aggregated or single-source?
  • Analytic methods → **Analytical Transparency**
Qualitative Research Transparency?

- Cultural opposition
- Incentives are lacking
- Human subjects/copyright concerns
- Where to put the material
- (Even) more work for researchers!
- “ Implicit” nature of qualitative methods
- No agreed transparency “techniques”
Qualitative Transparency Technique: Active Citation

**Traditional Format**

- Central or contested evidence-based claim in main text

+ Note with citation and supplementary text

**Transparency Appendix**

OVERVIEW ENTRIES, each with:

1. Full citation to source
2. Source excerpt
3. Annotation
4. (Potentially) link to the original source
Exercise – data for replication


Publishing_Exercise/Gambling_paper.pdf

• Skim read the research paper up to the Results section on Page 1674. Look at tables 1,2 and 3
• In groups: Identify what data (variables) and documentation you would definitely need to replicate the analysis
• Short class discussion
Issues to consider

• Which variables are needed; can these be grouped into groups?
• How many separate files might you need?
• What level of access might these data merit?

• Telephone survey – reliability; response rates
• Sampling strategy. List-assisted; one-plus
• Weighting to adjust for sample bias
• Specific definitions and exclusions from this subsample for the analysis. Exclusion thresholds selected e.g. (pathological gambling, depression). Definitions > 12 days in past year for alcohol, 5 days in past year for drugs, 2 weeks depression lifetime
• Lots of recoding from original data
Preparing a code book

• Start to compile a mini codebook for the variables you need, use word or Excel:

• Columns:
  • Variable names
  • Variable labels
  • Response codes
  • Code labels
Publishing data

• Create a metadata record – a record ‘about’ the data; upload your codebook

• Training ReShare repository: http://rd-training.essex.ac.uk/

• Log in using ‘rs03’, ‘rs04’, ‘rs05’,….. as both username and password

• If says can’t, click the advanced box then proceed
View data collection: Subset of Gambling Impact and Behavior Study, 1997-1999 to support published paper


This data is a subset of the Gambling Impact and Behavior Study telephone survey, a nationally representative sample of 2,417 adults, created to support the academic paper entitled, 'Health Correlates of Recreational Gambling in Older Adults' published in the American Journal of Psychiatry. In the analysis multivariate analyses were used to compare past-year recreational gamblers and non-gamblers in the older and younger age groups on measures of alcohol use and abuse/dependence, substance abuse/dependence, depression, mental health treatment, subjective general health, incarceration, and bankruptcy. Additional analyses compared the gambling patterns in older and younger adult past-year recreational gamblers.

Variables described in the analyses were derived from those described in the Gambling Impact and Behavior Study, with responses to questions grouped as indicated in the tables (Gerstein D, Hoffmann J, Larson C, Engelman L, Murphy S, Palmer A, Chucro L, Toce M, Johnson R, Bule T, Hill MA. Gambling Impact and Behavior Study. Chicago, University of Chicago, National Opinion Research Center, 1999). To obtain data on recreational gamblers, individuals with present or past problem or pathological gambling (N=51) were excluded using the criteria defined in the Gambling Impact and Behavior Study. As in the Gambling Impact and Behavior Study, respondents who reported three to four and five or more DSM-IV based diagnostic criteria for pathological gambling as assessed in the National Opinion Research Center diagnostic screen were categorized as being problem and pathological gamblers, respectively.

<table>
<thead>
<tr>
<th>Creators:</th>
<th>Creator Name</th>
<th>Email</th>
<th>Affiliation</th>
<th>ORCID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desai, R</td>
<td>Unspecified</td>
<td></td>
<td>Division of Substance Abuse of the Department of Psychiatry, Yale</td>
<td>Unspecified</td>
</tr>
</tbody>
</table>
Original codebook


- http://www.icpsr.umich.edu/cgi-bin/file?comp=none&study=2778&ds=1&file_id=909203&path=ICPSR